UNITED STATES
NAVAL POSTGRADUATE SCHOOL

THESIS

AN INTRODUCTION TO STATISTICAL DECISION FUNCTIONS

-by-

Lieutenant Ralph M. Tucker, U.S.N.

AN INTRODUCTION TO STATISTICAL DECISION FUNCTIONS

* * * *

Ralph M. Tucker

AN INTRODUCTION TO
STATISTICAL DECISION FUNCTIONS

by

Ralph McMath Tucker
Lieutenant, United States Navy

Submitted in partial fulfillment
of the requirements
for the degree of
MASTER OF SCIENCE

United States Naval Postgraduate School
Monterey, California

1955

This work is accepted as fulfilling

the thesis requirements for the degree of

MASTER OF SCIENCE

from the

United States Naval Postgraduate School

ABSTRACT

A non-technical discussion and the general technical formulation of a statistical decision problem are given. Following this, statistical decision theory is used to solve a testing problem concerning a prototype midget submarine. A set of rules to be followed in conducting the testing and reaching an optimum decision as to whether to accept the midget is developed. The development proceeds according to the Bayes solution of a statistical decision problem in which the stochastic variables are independently and identically distributed and limited to take only two values. Finally, brief discussions of the assumptions and restrictions of statistical decision theory and the role of the Minimax solution are included.

PREFACE

Operations research in the Navy is concerned with the establishment of quantitative basis for command decision. To help achieve this, the naval operations analyst is constantly seeking more useful tools. One such tool is the new theory of statistical decision functions, which, though presently unexploited in application, holds promise of extensive future use.

The theory of statistical decision functions was developed in the decade prior to 1950 by the late Abraham Wald. The development culminated in the publication of his definitive book Statistical Decision Functions . The book was written for mathematicians, and is too cryptic for the reader of limited mathematical background. This fact, along with the writer's belief that statistical decision theory can be of practical value to the naval operations analyst, prompted the present thesis. The thesis is intended as an introduction to the subject, and, except for Chapter I which is non-mathematical, is directed toward the reader who has studied calculus and has completed an elementary course in probability and statistics. The thesis purports to do no more than present the most essential elements of statistical decision theory and the detailed solution of a simple special case. The reader interested in a more mature treatment is referred to Wald.

Source material for the paper has consisted primarily of Wald's book and notes taken by the author during a course of instruction in statistical decision functions given by Professor Thomas E. Oberbeck at the United States Naval Postgraduate School. The contents are

arranged in five chapters and an appendix. Chapter I is a non-technical discussion of a type of practical problem that may be solved by statistical decision theory. The technical treatment begins in Chapter II, where the general formulation of the Bayes solution of the statistical decision problem is presented. Chapter III introduces certain assumptions needed to apply the theory, and Chapter IV treats an elementary special case. Chapter V deals with the Minimax solution. The Appendix gives a review of some selected mathematical concepts needed to understand better the technical discussions. It is recommended that the reader study the Appendix before beginning Chapter II.

The thesis was written during the period January - June , 1955 at the U. S. Naval Postgraduate School, Monterey, California. I wish to express my gratitude to the Navy for affording me the opportunity to write the thesis, to Professor Thomas E. Oberbeck for the technical competence and contagious enthusiasm he brought to his task as faculty advisor, to Professor Walter Jennings for helpful suggestions made while serving as second reader, and to Mrs. D. P. Slingerland for painstaking clerical assistance.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

TABLE OF SYMBOLS

(Listed in the order of their use in the text)

X	a stochastic process
$F(x)$	a joint cumulative probability distribution function
$G(x)$	a cumulative probability distribution function of one variable
$g(x)$	a probability density function of one variable
p	a parameter value
Ω	parameter space
P	a parameter viewed as a stochastic variable
$\xi(p)$	an a priori cumulative probability distribution function of P
$\xi'(p)$	an a priori probability density function of P
D^t	space of terminal decisions
D^e	space of decisions to continue experimentation
d^t	a terminal decision
d^e	a decision to continue experimentation
D	space of all decisions
W	weight function
C	cost function
s_k	the k^{th} stage of experimentation
δ	a statistical decision function
r	risk function
$\mathcal{D}$	space of statistical decision functions
R	a metric
$\{c_m\}$	a sequence
δ^m	a statistical decision function which prescribes no more than m observations

ρ_m	the least average risk considering only decision functions which prescribe no more than m observations
ρ	the least average risk
a	a value of X observed in a trial
$f^*(a\|p)$	the cumulative probability distribution function of a, given p
$f^*(a\|\xi)$	the expected cumulative probability distribution function of a, given ξ, the distribution of P.
c	the cost of one observation or trial
ξ_a	the a posteriori cumulative probability distribution function of P based on the observation a.
ξ_{ij}	the a posteriori cumulative probability distribution function of P after i 0's and j 1's have been observed
p_{ij}	the probability of obtaining the value 1 on a single observation when ξ_{ij} is the a priori cumulative probability distribution function of P

CHAPTER I

A NON-TECHNICAL DISCUSSION

1. Introduction.

In naval planning it is often necessary to predict the future usefulness of a proposed weapon or tactic. To do this, some value associated with the weapon or tactic, such as percentage success, average missed distance, or average life, is selected as a measure of the usefulness of the weapon or tactic. The problem then becomes one of estimating what this value, which we shall refer to as a parameter value, would be in a future war.

The usual procedure for doing this is to conduct some trials. An estimate of what the parameter value would be in a future war is obtained as a result of these trials. The important thing to note is that the estimate is not guaranteed to be correct. We intuitively suspect that the accuracy of the estimate increases as the number of trials conducted increases. Hence, the number of trials to be conducted is of fundamental importance.

The question of how many trials to conduct is often decided arbitrarily. Again, if the services of a statistician are available, the naval planner may determine the number of trials required to give, on the average, an arbitrarily specified degree of confidence in the estimate. In either case, some arbitrariness is retained.

Statistical decision theory adds a refinement to this procedure. It employs a criterion based on probability theory to select an optimum number of trials. The process involves a sort of cost analysis

of the problem. In practical situations the cost of conducting trials will usually be significant, and a definite cost may be associated with a poor estimate of the parameter value. To avoid the cost of the trials the planner is led to conduct no trials, or only a few; to avoid the cost of a poor estimate he is led to conduct a great number of trials. Obviously, the two considerations are opposed. The purpose of statistical decision theory is to reconcile these two opposing considerations, and, by the use of the criterion, to arrive at an optimum plan concerning the number of trials to be conducted and the final decision to be reached. Let us consider an example to see what this means.

2. Exhibit A.[1]

Suppose the Navy is interested in a newly developed midget submarine to be launched from a mother submarine and used to kill enemy submarines. The question of detection is not under consideration, but merely the capability of the midget to effect kills. It has been decided that the device should be tested. Budgetary considerations, considerations of priority of the services of the testing agency, etc. dictate the necessity of answering the question: How many trials are likely to be conducted? The question may be answered by using statistical decision theory. But before giving the answer, it is necessary to establish some precepts to be used in reaching it. The technical meaning of these precepts will be seen later, when we discuss each as a datum of the statistical decision problem. For the moment, let us think of them merely as the ingredients of a recipe. They must be put into the problem if we are to obtain a solution.

[1] The example is entirely fictitious (hence unclassified), and has been chosen merely for illustration.

There are five precepts, and they are straightforward. First, we decide to classify each trial of the midget submarine as a success or a failure, accordingly as the midget succeeds or fails to achieve a kill on the trial. Then, in our problem, the percentage success of the midget submarine in a future war is the unknown parameter value described in the Introduction. Second, we must say something about the relative likelihood of the various possible parameter values, i.e., the possible values of the percentage success of the midget submarine in a future war. Since we have no knowledge to the contrary, we assume that all values in the range 0 to 100% are equally likely to occur. Third, we decide to accept the midget if it succeeds on fifty percent or more of its trials, and to reject it if it does not. This decision might be based, for example, upon an assumption that present anti-submarine attack methods will succeed from twenty-five to seventy-five percent of the time in a future war. Fourth, we assume that the cost of each trial will be the same, and a study of the tactical situation, forces involved, etc. fixes the amount at $4000 per trial. Fifth, we have to establish the cost of a wrong decision as to whether the midget is superior to present anti-submarine attack methods. We may do this by making a careful study. The study might consider such things as the cost of producing the midget, the number that would be produced, the cost of alternative weapons, etc., and it leads us to an estimate of the cost of a wrong decision as shown in the following table:

(See following page for table).

Decision	Percentage success of midget in future war	Cost in Dollars
Accept midget (trial successes greater than 50%)	more than 25%	0.
Accept midget (trial successes greater than 50%)	less than 25%	1,000,000.
Reject midget (trial successes less than 50%)	more than 75%	1,000,000.
Reject midget (trial successes less than 50%)	less than 75%	0.

Cost of Decision

Table 1.

Note that the first and last lines of the table represent correct decisions and cost nothing, whereas the second and third lines represent wrong decisions and cost a definite amount.

Let us elaborate upon the nature of these "costs" of wrong decision. They are not actual amounts of money that must be paid to someone. Rather, they may be explained as follows: If a wrong desion is made, a certain disadvantage accrues to the Navy as a result. The money evaluation of this disadvantage is called the cost of wrong decision. This is much like the "cost" to a salesman who loses a $300 commission because he elects to play golf instead of seeing a prospective customer. He doesn't have to pay anyone the $300, but he is nonetheless $300 worse off than he might have been. We say he has made a "costly decision". In short, the cost of wrong decision

is the money equivalent of the loss suffered as a result of the wrong decision.

In considering the costs shown in Table 1 it is necessary to keep in mind the distinction between the percentage success observed in the trials and the percentage success the midget submarine would have in a future war. The first is an estimate of the second, and is not necessarily correct. Note that, if the midget is definitely superior to present anti-submarine attack methods (i.e., would have a percentage success in a future war in excess of 75%) and we reject it, we must penalize ourselves $1,000,000. This is the cost shown in the third line of the table. Similarly, if the midget is definitely inferior to present anti-submarine attack methods (would have a percentage success in a future war of less than 25%) and we accept it, we must again penalize ourselves $1,000,000. This is the cost shown in the second line of the table. On the other hand, if the percentage success of the midget in a future war is between 25% and 75%, we do not need to penalize ourselves for either decision. This is not unreasonable in view of our earlier assumption that present anti-submarine attack methods have a percentage success of 25% to 75%. For, if the midget would have a percentage success in the same range, we shall consider that we have really neither gained nor lost by either accepting or rejecting it.

The solution to the problem can now be given. It takes the form of a table (Table 2). The question of how such a table is obtained is, essentially, the subject of this paper. The actual detailed procedure for obtaining this particular table is presented in Chapter IV. For

the moment, let us accept the table. We can then examine and interpret it, so that we may gain an appreciation of the role of statistical decision theory.

(See following page for Table 2.)

Let us note the construction of the Table. The numbers identifying the rows and columns designate, respectively, the number of failures and successes of the midget submarine that have been observed in successive trials. Any set of one row designator and one column designator locates a square in the table. This square then applies to the situation existing after the indicated number of failures and successes have been observed. For example, the square in row number three and column number five applies after three failures and five successes have been observed. Each square contains two numbers (of dollars). They have the following meanings:

upper number: the anticipated cost (in dollars) to the Navy if no further trials are conducted, and a decision is made to accept or reject the midget on the basis of the trials conducted thus far.

lower number: the anticipated cost (in dollars) to the Navy if trials are continued, and a final decision to accept or reject the midget is based on the results of further trials.

The choice of the words "anticipated cost" in defining these two numbers has been carefully made. This is because the dollar values represented by these numbers in the table are "expected values" in the sense of probability theory. This is discussed in the Appendix. What the reader must understand at this point is

NO. OF SUCCESSES OBSERVED

NO. OF FAILURES OBSERVED

	0	1	2	3	4	5	6	7	8	9	10
0	250,000 25,200	62,500 21,200	15,600 12,600	3,900 3,900	1,000 1,000	200 200	100 100	0 0	0 0	0 0	0 0
1	62,500 21,200	156,300 26,500	50,800 22,500	15,600 14,100	4,600 4,600	1,300 1,300	400 400	100 100	0 0	0 0	0 0
2	15,600 12,600	50,800 22,500	103,500 25,000	37,600 21,000	12,900 12,900	4,200 4,200	1,300 1,300	500 500	100 100	0 0	0 0
3	3,900 3,900	15,600 14,100	37,600 21,000	70,600 22,500	27,300 18,500	10,000 10,000	3,500 3,500	1,200 1,200	400 400	100 100	0 0
4	1,000 1,000	4,600 4,600	12,900 12,900	27,300 18,500	48,900 21,000	19,800 16,100	7,600 7,600	2,800 2,800	1,000 1,000	300 300	100 100
5	200 200	1,300 1,300	4,200 4,200	10,000 10,000	19,800 16,100	34,300 17,600	14,300 13,600	5,600 5,600	2,200 2,200	800 800	300 300
6	100 100	400 400	1,300 1,300	3,500 3,500	7,600 7,600	14,300 13,600	24,300 14,300	10,300 10,300	4,200 4,200	1,600 1,600	600 600
7	0 0	100 100	500 500	1,200 1,200	2,800 2,800	5,600 5,600	10,300 10,300	17,300 11,500	7,500 7,500	3,100 3,100	1,200 1,200
8	0 0	0 0	100 100	400 400	1,000 1,000	2,200 2,200	4,200 4,200	7,500 7,500	12,400 9,400	5,400 5,400	2,300 2,300
9	0 0	0 0	0 0	100 100	300 300	800 800	1,600 1,600	3,100 3,100	5,400 5,900	9,000 7,900	3,900 3,900
10	0 0	0 0	0 0	0 0	100 100	300 300	600 600	1,200 1,200	2,300 2,300	3,900 3,900	6,900 6,900

Solution of Exhibit A

Table 2.

that the numbers are not absolute like the $4000 cost per trial and the $1,000,000 cost of wrong decision. Rather, they are values calculated on the basis of the likelihoods of occurrence of the possible outcomes, much like insurance companies calculate the life expectancy of man from the relative frequency of deaths at each age.

The anticipated costs shown in Table 2 constitute the criterion used to determine an optimum solution to the problem. Hence, the solution is optimum relative to these costs as a criterion. Since the nature of the criterion is probabilistic, the final decision, also, is probabilistic. What this means, in practical terms, is that, if a rare and unlikely series of results is obtained on the conducted trials, such as success on every trial when actual future wartime employment will yield a preponderance of failures, a poor decision will be made. This is a chance that must be taken to avoid the great cost that would certainly occur if a very large number of trials were conducted. It does not invalidate the theory any more than the survival of one individual to age 106 invalidates the methods of insurance companies.

We may now proceed with the interpretation of the table. Notice that the upper number is greater than the lower number in some of the squares, and equal to it in others. Those in which it is greater are enclosed within the double lines. At any stage of testing corresponding to one of these enclosed squares, the anticipated cost is less to continue taking trials than it is to reach a decision at that time. On the other hand, at any stage of testing corresponding to a square outside the double lines, the anticipated cost is as little if trials are halted, and a decision to accept or reject the midget submarine is made

on the basis of the sample already taken.

Now observe that the (0, 0) position is within the double lines. This means that the initial anticipated cost is least if some trials are conducted. The number that will be conducted depends on the outcome of the trials. We begin in the (0, 0) position, and conduct a trial. If it succeeds, we move right to the (0, 1) position; if it fails, we move down to the (1, 0) position. In either case, the second position is still within the double lines, so another trial is conducted. This process is continued until a position outside the double lines is reached. This may require anywhere from three to 13 trials. For example, if the first three trials all succeed, position (0, 3) will be reached. Here, the upper entry ceases to be greater than the lower entry, so it will pay to stop taking trials and decide, since the percentage success of the trials conducted of 100% is greater than 50%, to accept the midget. As another example, suppose the trial outcomes alternate from success to failure to success, etc., in that order. This will result in a stair stepping down the table, returning to the main diagonal (number of successes equal number of failures) on alternate trials. Eventually we must arrive outside the double lines in position (6, 7) after 13 trials. The percentage success for the trials conducted is then

$$\frac{7}{13} \times 100 = 53.8\% ,$$

and again the decision is made to accept the midget. If the sequence of outcomes leads to a position outside the double lines on the upper side of the main diagonal, the percentage success of the trials

conducted will be greater than 50%, and the midget will be accepted; if it leads to a position outside the double lines on the lower side of the main diagonal, the percentage success of the trials conducted will be less than 50%, and the midget will be rejected.

With the aid of Table 2, it is now possible to answer the earlier question of how many trials are likely to be conducted. The answer consists of Table 2 and the following rule:

> Begin conducting trials, and following each trial, note the position reached in the table. Continue this until a position outside the double lines is reached, then accept the midget if the number of successes exceeds the number of failures. Reject the midget if the number of failures exceeds the number of successes. The minimum number of trials required to reach a final decision will be three; the maximum number will be 13.

3. Another Aspect.

A direct solution of the problem has been given. Let us now consider a possible budgetary complication. Suppose that $32,000 has been alloted to conduct the testing of the midget submarine. This is, of course, an illogical amount in the light of statistical decision theory. The solution does not divulge exactly how much the testing will cost. It predicts only that from three to 13 trials will be required. At $4000 per trial, this amounts to a cost of from $12,000 to $52,000. The dilemna can only be resolved, in the light of statistical decision theory, by getting the allotment changed to permit the flexibility required by the solution. Failing this, an optimal decision may be reached, but cannot be guaranteed. If it turns out that a position

outside the double lines, such as (5,3), is reached within eight trials, an optimum decision will be reached in spite of the limitation. On the other hand, if we are still inside the double lines after eight trials, such as in position (4,4), sufficient data to indicate an optimum decision has not been collected.

A variation of the budget problem is the case in which more than $52,000 is available for the testing. In such a case, expenditure beyond $52,000 is, according to statistical decision theory, a waste of funds. The solution will have indicated the optimum decision after 13 trials, if not before, and additional trials are not called for by the theory.

4. Summary.

Exhibit A has been studied to help provide a conceptual understanding of what is involved in the type of solution of the testing problem provided by statistical decision theory. It should be remembered that the precepts, i.e., the decision to classify each trial as a success or a failure, the decision to either accept or reject the midget, the specification of the cost of testing, the specification of the cost of wrong decision, and the specification of the likelihood of various values of the percentage success in a future war, are necessary inputs to the problem. Finally, the solution that is obtained is optimum relative to the anticipated costs as the criterion, and the final decision is probabilistic in nature.

CHAPTER II

GENERAL FORMULATION OF THE BAYES SOLUTION

1. Basis of the Problem.

A datum of any problem is defined to be something, actual or assumed, that is used as a basis for reckoning. The statistical decision problem has five of these. In Exhibit A we considered them intuitively as precepts. Let us now examine them in more technical detail, and introduce a portion of the notation of statistical decision theory.

a. Stochastic Process X: A stochastic process is defined as a countable collection of stochastic (chance) variables having a joint cumulative probability distribution. To explore this, let us think of a countable collection of stochastic variables

$$X = \{X_i\} = \{X_1, X_2, X_3, \ldots\ldots\}.$$

Let us next think of a countable set of real values, one for each stochastic variable, i.e.,

$$\underset{\sim}{x} = \{x_i\} = \{x_1, x_2, x_3, \ldots\ldots\}.$$

By definition (see Appendix), the joint cumulative probability distribution of all the stochastic variables in the countable collection is the probability that $X_i \leq x_i$ simultaneously for all i. In other words, $F(x)$ is the probability that $X_1 \leq x_1$, $X_2 \leq x_2$, $X_3 \leq x_3$, ... simultaneously.

An important special case of a stochastic process should be mentioned. It is the case where the stochastic variables X_1, X_2,

. . . . . are independently and identically distributed. The condition of independence means that the joint distribution function is the product of the individual distribution functions. In this case,

$$F(x) = G_1(x_1)\, G_2(x_2)\, G_3(x_3) \ldots\ldots = \prod_{i=1}^{\infty} G_i(x_i),$$

where $G_i(x_i)$ is the distribution function of the i^{th} stochastic variable. The condition that the stochastic variables be identically distributed means that the distribution of each stochastic variable has, not only the same form (such as normal or uniform), but also the same parameter values. Thus, we might have

$$G_i(x_i) \equiv G(x_i;\mu,\sigma) \qquad \text{for all } i$$

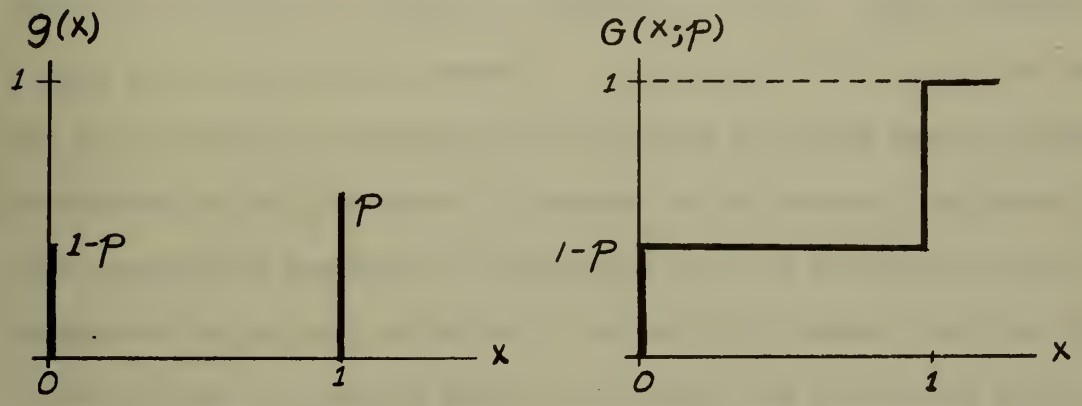
where $G(x;\mu,\sigma)$ denotes a normal distribution with mean μ and standard deviation σ. In this case, we may write

$$F(x) \equiv F(x;\mu,\sigma) = \prod_{i=1}^{\infty} G(x_i;\mu,\sigma)$$

showing the dependence of F upon the values of the parameters, μ and σ.

The stochastic process of Exhibit A is an example of one in which the stochastic variables, X_i, are assumed to be independently and identically distributed. The outcome of each trial of the midget submarine is considered to be a stochastic variable. Hence, the result of the i^{th} trial constitutes the stochastic variable X_i. The possible particular outcomes of each trial, success or failure, are thought of as representing particular values of the stochastic variables. The assumption that every trial has the same chance of succeeding as every other trial is equivalent to the assumption that

the stochastic variables are identically distributed. The two values to which the stochastic variables are restricted (failure and success) are denoted by 0 and 1 respectively. The common percentage success of each stochastic variable, thought of as a parameter, is labeled p (parameter value). This makes it possible to depict the stochastic process diagramatically by showing the distribution, G (x;p), of one of the identically distributed stochastic variables as in Figure 1.



Bar Graph Step Function

Distribution of One of the Stochastic Variables of Exhibit A

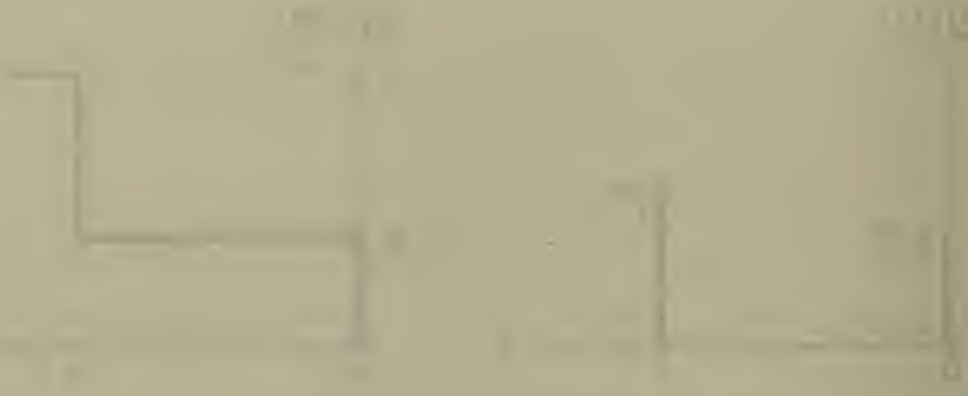
Figure 1.

In this case, we may write

$$F(x) = F(x;p) = \prod_{i=1}^{\infty} G(x_i;p)$$

showing the dependence of F upon the parameter p.

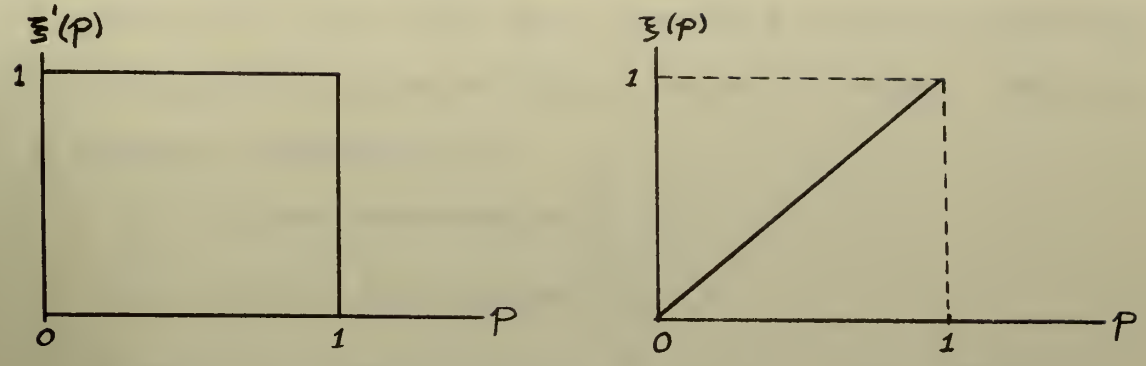
b. Space Ω : The space Ω is defined to be a class of joint cumulative probability distribution functions known to contain the true distribution, F (x), of the stochastic process. The elements of Ω are joint cumulative probability distribution functions and differ from one another only in the values of their parameters. Hence, $F(x;p_1)$, $F(x;p_2)$, ... , $F(x;p_n)$, . . . are elements of the space Ω



when F depends on a single parameter. For this reason, it is often convenient, as well as illuminating, to think of Ω as a parameter space. Adopting this view in subsequent portions of this paper, we shall refer to elements of the space Ω as values of this parameter. The parameter is then regarded as a stochastic variable, P, and, as such, is liable to take on different values with different likelihoods. Note that, following convention, we denote the parameter in its role as a stochastic variable by using the capital letter P, while parameter values are denoted by the small p. In short, Ω is a class of similar joint cumulative probability distribution functions having different parameter values and known to contain, as an element, the particular joint cumulative probability distribution function having the correct parameter value, or, as we have referred to it above, the true F. To determine an optimum way of estimating this parameter value is the crux of the statistical decision problem.

In Exhibit A, the percentage success of the midget submarine in a future war is the particular parameter value of interest. If we knew it, there would be no problem. Since we do not, we regard the unknown parameter as a stochastic variable, P. We know only that this stochastic variable is confined to range between 0 and 100%. It was stated, as a precept, that prior to any experimentation we would assume the true parameter value to be anywhere in this range with equal likelihood. This is equivalent to saying that the stochastic variable, P, is continous and that its a priori distribution is uniform. The uniform probability density function of P, $\xi'(p)$, and the associated cumulative probability distribution

function, $\xi(p)$, are shown in Figure 2.



A Priori Distribution of the Parameter of Exhibit A

Figure 2.

Note that p represents a possible value of the stochastic variable P, and $\xi(p)$ represents the probability that $P \leq p$.

c. Space D^t: The space D^t is defined to be the space of possible final decisions. To illustrate D^t, let us again refer to Exhibit A. We recall that, at any stage of experimentation, we were always faced with two alternative types of decisions, namely, to make a final decision or to continue experimenting. These two types of decisions are distinguished by defining two classes of decisions:

D^t: the class of all terminal decisions

D^e: the class of all decisions to continue experimenting, such as take one more trial or take two more stages of three trials each, etc.

Now, in Exhibit A, D^t consisted of two elements:

d_1^t: accept the midget.

d_2^t: reject the midget.

D^e consisted of a single element:

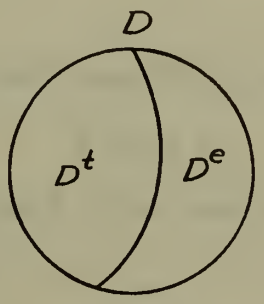
d_1^e : take one more trial.

In general, D^t and D^e are not so restricted, but may consist of as many elements as needed to cover all possible decisions. This idea is expressed symbolically by:

D^t is a class consisting of d_1^t , d_2^t , ...

D^e is a class consisting of d_1^e , d_2^e , ...

To illustrate the relation between D^t and D^e , it is convenient to define the class D as the class of all possible decisions. It is then clear that $D = D^t \cup D^e$. This is shown pictorially in Figure 3.



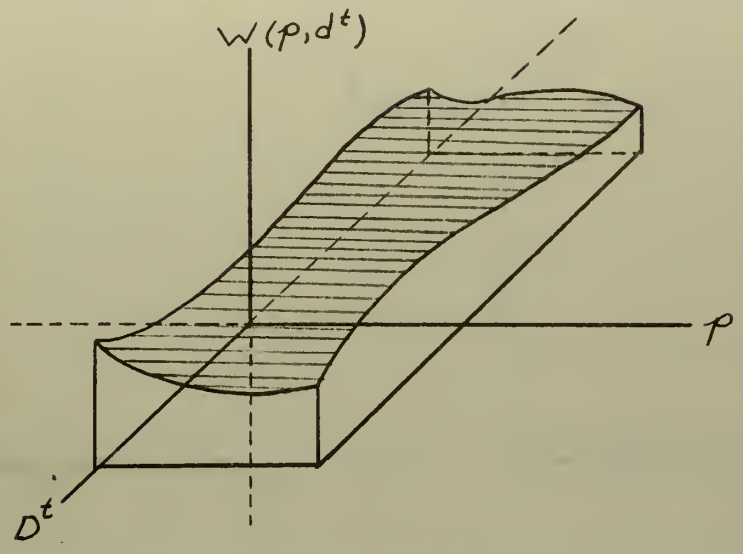
Decision Space

Figure 3

It will be recognized that the sum total of all decisions from D^t and D^e are exhaustive and mutually exclusive.

d. Weight Function $W(p, d^t)$: The weight function is defined to be a non-negative function, the value of which expresses the cost of making the terminal decision d^t when the true parameter value is p . It is through the weight function that the cost of making a wrong decision is introduced into the problem. If a correct decision is made, the value of $W(p, d^t)$ will be zero; if an incorrect decision is made, $W(p, d^t)$ may have a positive value. In general, the

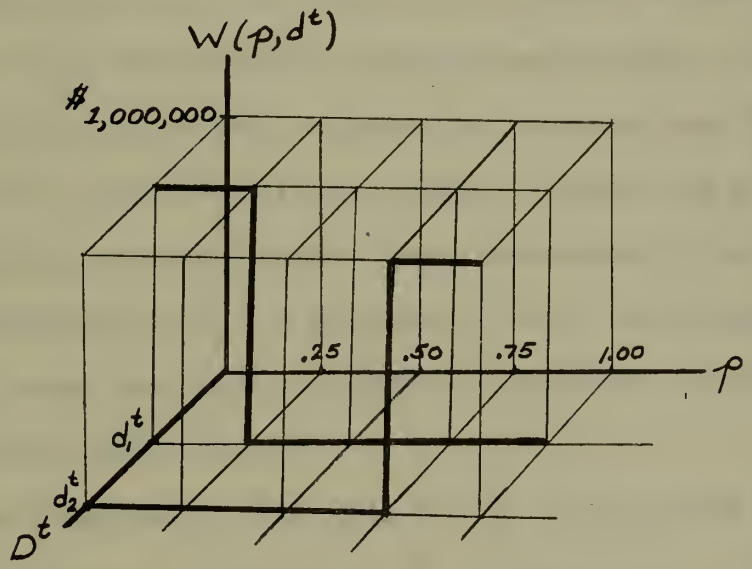
weight function, like any function of two variables, may be depicted as a surface as shown in Figure 4.



Weight Function (General)

Figure 4.

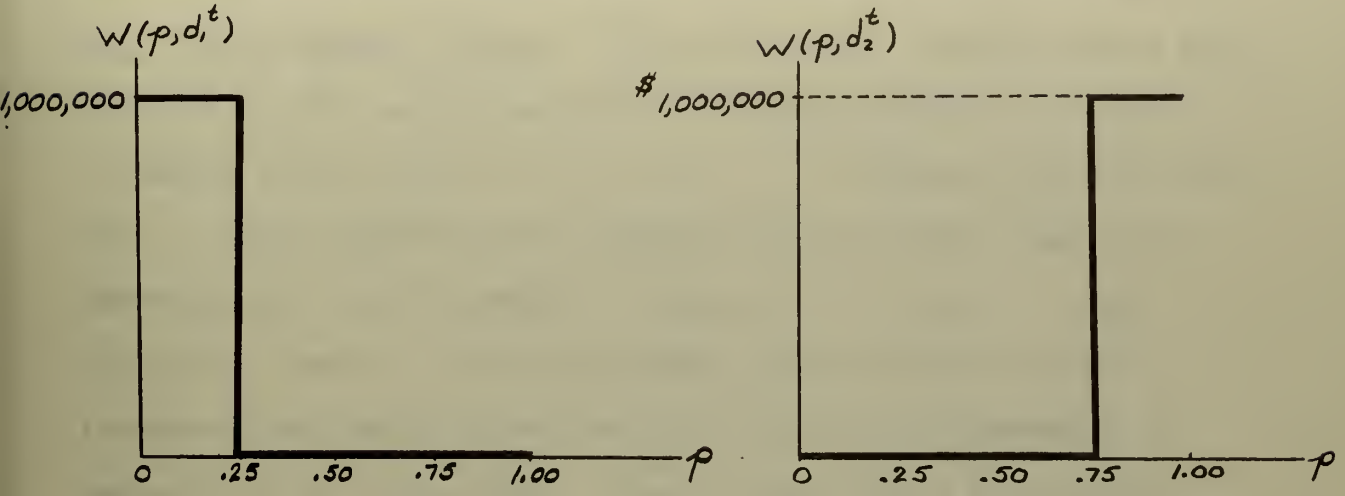
In the special case of Exhibit A, this surface degenerates into two curves in space as follows:



Weight Function for Exhibit A

Figure 5.

Since there are only two elements in D^t for Exhibit A, it becomes more instructive to represent Figure 5 as two curves as shown in Figure 6.



Alternative Representation of Weight Function for Exhibit A

Figure 6.

Either Figure 5 or Figure 6 is the equivalent of Table 1.

The weight function is the most difficult datum of the statistical decision problem to specify. Since it is a datum, it must be known before a statistical decision problem can be solved. The operations analyst must be able to specify its value for any values of the arguments p and d^t. This amounts to saying that he must be able to assign a numerical cost to any combination of a possible terminal decision and a possible parameter value. The question of how to do this is one that needs extensive investigation, and offers an opportunity for further study.

It is often possible and desirable to classify decisions as merely right or wrong. In such case, the weight function, $W(p, d^t)$, takes on only values 1 and 0, and is said to be a simple weight

function. Except for a scaling factor of 10^6 , this was the case in Exhibit A.

e. Cost Function C (x, s) : The cost function is defined to be a non-negative function expressing the cost of experimentation. In general, it depends on the values, $x = x_1, x_2, \ldots$, obtained on the observations. It also depends on the variables observed in each stage of experimentation, and the number of stages, $s = s_1, s_2, \ldots s_k$, observed. However, it may be possible, and is usually desirable, to consider the special case in which the cost of experimentation is the same for each experiment. Then the total cost of experimentation is proportional to the number of trials conducted. This was the case in Exhibit A where each observation cost $4000, and the cost function had a value of 4000 times the number of observations taken.

2. The Statistical Decision Function, $\delta(x, s)$.

A statistical decision function, δ , is a set of rules which estimates a parameter using the results of observations of a stochastic process X. It depends on the values $x = x_1, x_2 \ldots$ obtained on the observations and on the variables observed in each stage of experimentation as well as the number of stages, $s = s_1, s_2, \ldots s_k$. It is a function which prescribes a plan for conducting experimentation and reaching a terminal decision. For example, in Exhibit A the statistical decision function consisted of the Table 2, from which instructions for experimenting and reaching a terminal decision were obtained. The problem of statistical decision theory is, given the stochastic process X , the space Ω , the space D^t ,

the weight function $W(p, d^t)$ and the cost function $C(x, s)$, to find the statistical decision function that provides the optimum decision.

3. The Risk Function, $r(p, \delta)$.

Each statistical decision function δ is an element of the class $\mathcal{D}$ of all statistical decision functions. To select that δ from $\mathcal{D}$ which provides the optimum solution to a statistical decision problem, a criterion is needed. That is the role of the risk function. We have already seen, from Exhibit A, that the criterion must take account of the conflicting costs of experimentation and wrong decision. To introduce these costs more precisely into the risk function, let us define

$r_1(p, \delta)$: the expected cost of decision [expected value of $W(p, d^t)$] when p is true and δ is used.

$r_2(p, \delta)$: the expected cost of experimentation [expected value of $C(x, s)$] when p is true and δ is used.

Note that $r_1(p, \delta)$ and $r_2(p, \delta)$ are both expected values. The meaning of "expected value" has been discussed briefly in connection with the anticipated costs of Exhibit A; it is explained more technically in the Appendix. Now, $r_1(p, \delta)$ and $r_2(p, \delta)$ are, respectively, the expected (average) values of $W(p, d^t)$ and $C(x, s)$ for given values of p and δ. That is, $W(p, d^t)$ is averaged by the probability that d^t will be made to give $r_1(p, \delta)$, and $C(x, s)$ is averaged by the probability that the values x will be obtained when the stages, $s = s_1, s_2, \ldots . s_k$, are observed to give $r_2(p, \delta)$. The notion that these averages are obtained for a particular (p, δ) should be kept clearly in mind, for we shall subsequently

require expected values calculated with respect to the variables p and δ. The risk function may now be defined to be the sum of the expected values of the weight function and the cost function for given values of p and δ. That is,

$$r(p, \delta) = r_1(p, \delta) + r_2(p, \delta).$$

Hence, the risk function, which may take on a value for any pair of arguments (p, δ), represents the total expected cost associated with these arguments.

4. The Bayes Solution.

The goal of statistical decision theory is to select the particular statistical decision function, δ_o, that prescribes the optimum plan concerning the number of trials to be conducted and the optimum terminal decision based on the results of these trials. The risk function is the basic criterion to be used in making this selection. But the risk function, as we have seen, depends on both p and δ for its value. The dependence on p makes it unsuitable, in its present form, as a yardstick for comparing the relative merits of various δ. To overcome this difficulty, we need to remove the dependency on p. This is accomplished by averaging out the p, leaving a new function, the average risk, which depends on δ alone for its value. The values of the new function may be ordered as to magnitude, and the magnitudes will vary with δ alone.

Let us elaborate on this. It often happens that a reasonable estimate of the likelihood of P taking on various values, p, can be given at the outset. That is, the physics of the problem, a study of past results, or even a shrewd analysis may provide an a priori distribution of P. This simply means that we are able to specify, either as an assumption or

as a reasonable approximation, some distribution function, $\xi(p)$, that describes the likelihood with which P will take on the values within its range of possible values. From this point on, ξ is assumed to be known.[2] If we now take the expected value of the risk function with respect to the a priori distribution of P, we get, in Wald's notation,

$$r(\xi, \delta) = \int_{\Omega} r(p, \delta) d\xi$$

Notice that this average risk, averaged with respect to the a priori knowledge of P, depends only on ξ and δ, and ξ is known. This is a significant result. It means that the average risk is suitable as the yardstick for comparing δ, since it can be ordered as to magnitude, and the magnitudes depend only on δ. Our interest, of course, is in selecting a particular δ_o that makes the average risk the least. That is, we want a δ_o such that

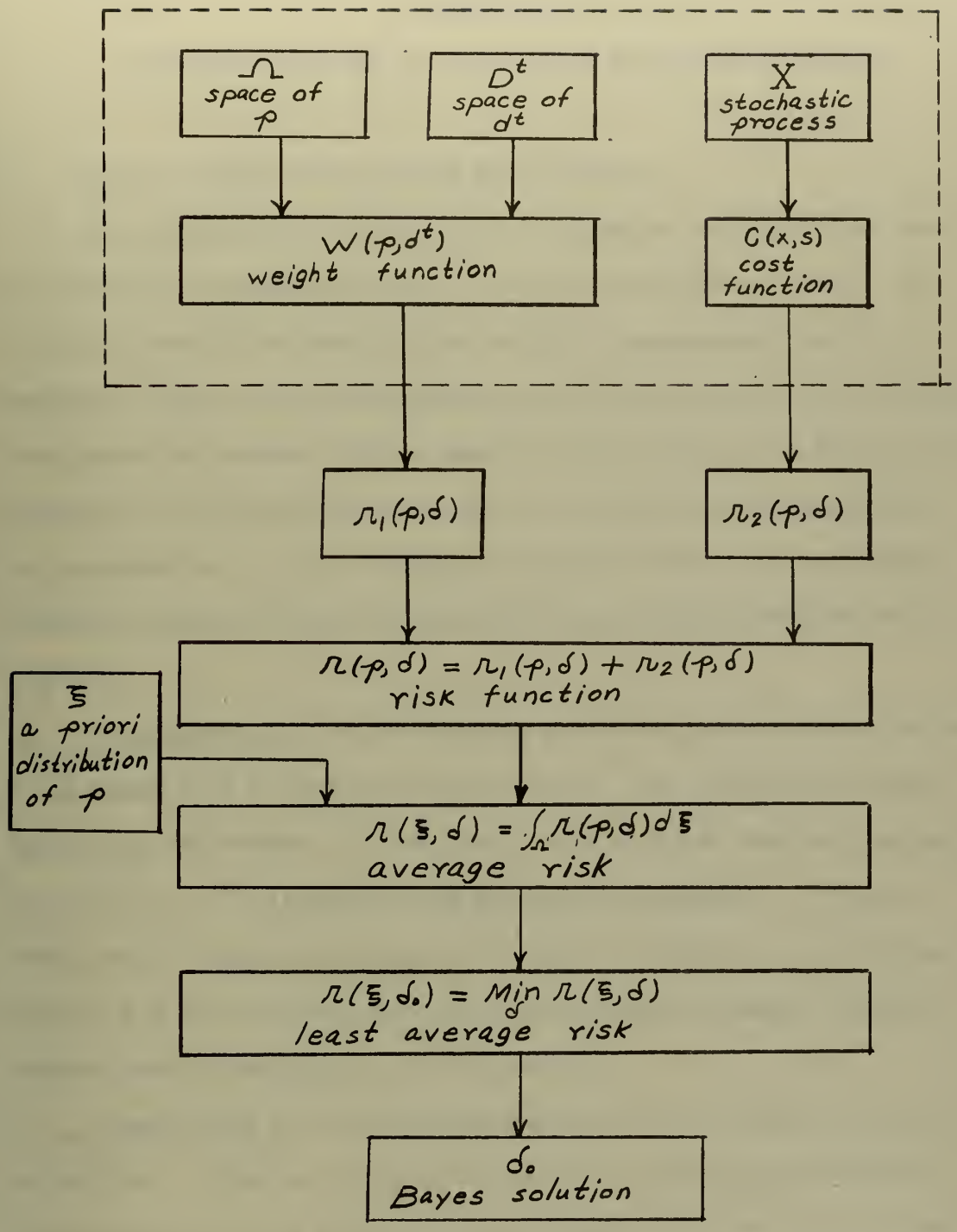
$$r(\xi, \delta_o) = \min_{\delta} r(\xi, \delta).$$

This is often alternatively expressed as

$$r(\xi, \delta_o) \leq r(\xi, \delta) \text{ for all } \delta.$$

Such a δ_o constitutes a Bayes solution. Thus, a Bayes solution is a δ_o which minimizes the average risk, $r(\xi, \delta)$, with respect to all δ. It is to be noted that a Bayes solution is a solution relative to a particular a priori distribution ξ. The procedure employed to arrive at a Bayes solution is summarized in the block diagram of Figure 7.

[2] The case where ξ cannot be specified is discussed in Chapter V.



Block Diagram of the Buildup to a Bayes Solution

Figure 7.

CHAPTER III

ASSUMPTIONS OF STATISTICAL DECISION THEORY

1. An Assumption Concerning Each Datum.

This chapter introduces some assumptions applied to the theory of statistical decision functions to insure that solutions exist. A complete study of the implications of these assumptions is not attempted. Rather, the assumptions are briefly presented here merely to acquaint the reader with the nature of the problem, so that he may gain some insight into the character of the restrictions imposed by the assumptions. A full treatment is given by Wald. One assumption regarding each datum of the statistical decision problem is required.

a. Assumption 1: The assumption regarding the stochastic process X is stated only for the case where the X_i are independently and identically distributed. In this case, it is assumed that the stochastic process, X, is discrete or absolutely continuous. That is, either each component stochastic variable is discrete, or it is continuous and has a density function. Continuous stochastic variables without density functions are not admitted.

b. Assumption 2: A convergence property regarding the space Ω is required. However, it is not necessary to explore the nature of this property for our purposes, since Wald shows that it is a consequence of Assumption 1. As such, it constitutes no additional limitation.

c. Assumption 3: The weight function, $W(p, d^t)$, is a bounded

function of p and d^t. Recalling that the weight function was defined to be a non-negative function which describes the cost of making any particular terminal decision, d^t, we see that this assumption merely excludes the possibility of any decision costing an infinite amount.

d. Assumption 4: The space D^t is compact in the sense of the metric

$$R(d_1^t, d_2^t) = \operatorname{Sup}_p \left| W(p, d_1^t) - W(p, d_2^t) \right| .$$

This assumption is fulfilled if the space D^t is finite. That is, if the number of terminal decisions which may be made is finite, the assumption is satisfied. This will cover most cases. However, if D^t is not finite, the assumption can generally be satisfied by restricting the range of the unknown parameter to a bounded space. This restriction appears to present no practical difficulty.

e. Assumption 5: The cost function, $C(x,s)$, satisfies the following three conditions:

(1) $C(x,s) \geq 0$ for all x and s, and $C(x;s_1, \ldots . s_{k+1}) \geq C(x;s_1, \ldots . s_k)$.

(2) For any given s, the cost, $C(x,s)$, is either a bounded function of x or $C(x,s) = \infty$ identically in x.

(3) There exists a sequence, $[c_m]$, $(m = 1,2, \ldots, \text{ad. inf.})$ of positive values such that

$$\operatorname{Lim}_{m=\infty} c_m = \infty , \text{ and}$$

$C(x,s) \geq c_m$ for all x, and for all $s = [s_1, \ldots, s_k]$ for which the set theoretical sum of $s_1, \ldots, s_k$ contains at least m elements.

The meaning of this assumption concerning the cost of experimentation is given in words as follows:

(1) The cost of experimentation cannot be negative, and the total cost of experimentation after an additional stage is taken cannot be less than it was before.

(2) The cost of experimentation is either finite or it is impossible to make observations of certain variables.

(3) Regardless of the values of the observations made or the number of stages employed in making them, if the total number of observations is at least m, then the cost, $C(x,s)$, of these observations is not less than the m^{th} term in some increasing sequence, c_m, which approaches infinity as a limit. The basic idea of this is that there exists some minimum value of the cost of observing m variables beyond which it is impossible to reduce the cost of observing m variables by rearranging the composition of the stages of experimentation. In other words, it is not possible to observe more variables for less money by taking the stages wholesale.

2. An Assumption Concerning the Space $\mathcal{D}$.

An assumption concerning the space $\mathcal{D}$ of admissible decision functions is made in addition to the assumptions concerning each datum. The most essential portion of the assumption is that only those decision functions which prescribe a finite amount of experimentation and which lead to a terminal decision are to be considered.

3. Some Consequences of the Assumptions.

Regardless of how slight the cost of experimentation, if one experimented an infinite amount the cost would increase without bound. Therefore, there exists a point beyond which further experimentation

is not profitable. This intuitive notion is developed rigorously by Wald when he shows that, even though we limit ourselves to decision functions which prescribe a finite amount of experimentation, we can still approach an optimum solution arbitrarily closely under the assumptions of this chapter.

Subject to the assumptions of this chapter, a Bayes solution exists for any given a priori distribution, ξ . If it is not practicable to specify an a priori distribution, then the decision problem may be viewed as a zero-sum, two person game in the sense of von Neumann's theory of games, and a minimax solution exists. A minimax solution is a Bayes solution relative to the least favorable a priori distribution. The minimax solution is discussed further in Chapter V .

CHAPTER IV

THE BAYES SOLUTION FOR A SPECIAL CASE

1. General.

The general formulation of the Bayes solution to the statistical decision problem was given in Chapter II, and some of the theory underlying its development was pointed out in Chapter III. In this chapter, we shall undertake a progressive restriction of the general problem until, ultimately, we arrive at the special case illustrated by Exhibit A. Thereupon, the detailed solution of Exhibit A will be indicated. The first step in this process will be to consider a statistical decision problem in which the stochastic variables are restricted to be independently and identically distributed, and the cost of experimentation to be proportional to the number of observations. Then we shall proceed to the case where the stochastic variables are further restricted to take only two values. The discussion of the latter will terminate with the solution of Exhibit A.

2. Independently and Identically Distributed Stochastic Variables with Simple Cost.

Recalling that the object of statistical decision theory is to find the "best" decision function, we may readily see how the restrictions we are imposing will help us. By restricting the cost function to be simple, i.e., by requiring the cost of experimentation to be proportional to the number of observations, we make it possible to ignore the manner in which the observations are grouped or arranged. That is, we may consider only those decision functions for which

each stage of experimentation consists of exactly one observation. Further, by requiring the stochastic variables X_i to be independently and identically distributed, we eliminate the need for concern as to which particular stochastic variables are observed. As a consequence, we may limit the decision functions considered to those which not only prescribe a single observation per stage, but also prescribe that the stochastic variables will be observed in order. This is possible because the stochastic variables, being identical, may be ordered in any desired way.

In continuing our search for a "best" decision function, we may now assert that, in choosing it, we need only compare the merits of decision functions falling into the limited category explained in the preceding paragraph. And since we are seeking a Bayes solution, the decision function we ultimately select will be the one that is "best" in the sense of the Bayes solution of Chapter II. The reader will recall that the Bayes solution is given relative to an a priori distribution $\xi(p)$ in Ω, and that it consists of that decision function, δ_0, which minimizes the average risk - the average being taken with respect to ξ and the minimum over all δ. With these facts in mind, we may proceed with the process of comparing the average risk produced by each δ, and the choice of the δ_0 which produces the least average risk.

Let m be a non-negative integer, and let δ^m denote a decision function which guarantees that the total number of observations will not exceed m. Then, for any a priori distribution ξ, we may define

$$\rho_m(\xi) = \operatorname{Min}_{\delta^m} r(\xi, \delta^m)$$

to be the least average risk that can be found by considering only decision functions which guarantee no more than m observations. Similarly,

$$\rho(\xi) = \operatorname{Min}_{\delta} r(\xi, \delta)$$

is the least average risk to be found by considering all decision functions, whether or not they prescribe a finite number of observations. A particular decision function that belongs to both classes δ and δ^m, which we will be interested in, is δ^o. This is the decision function which is guaranteed to prescribe no observations. It is of interest because it enables us to write

$$\rho_o(\xi) = \operatorname{Min}_{\delta^o} r(\xi, \delta^o) = \operatorname{Min}_{d^t} W(\xi, d^t) .$$

This is an obvious, but important relation. It says simply that the least average risk, if we consider only decision functions which prescribe no experimentation, is equal to the minimum cost of decision. This follows from the definition of risk (cost of experimentation plus cost of decision) as given in Chapter II, and the fact that no experimentation is involved.

Two remarks at this point may assist the reader in avoiding misunderstanding. First, whereas cumulative distribution functions (such as Ξ) are usually employed in logical developments, the corresponding density functions (such as Ξ') are more often used in calculation. The distinction should be constantly remembered. Second, the present chapter requires Assumptions 1-5 of Chapter III, but does not require the assumption concerning $\mathcal{D}$ - a fact the reader may have surmised from the introduction of $\rho(\xi)$.

There are several theorems concerning the functions $\rho_m(\xi)$, $\rho(\xi)$ and $\rho_o(\xi)$ which enable us to compare various average risks and lead us to the Bayes solution. Perhaps the most important of these is the recursion formula

$$(A) \qquad \rho_{m+1} = \operatorname{Min} \left[\begin{matrix} \rho_o(\xi) \\ c + \int_{-\infty}^{\infty} \rho_m(\xi_a)\, df^*(a|\xi) \end{matrix} \right]$$

We need to examine this formula carefully and understand it thoroughly. It contains several symbols not given explicitly before. They are

a: stands for a value that might be obtained if a stochastic variable were to be observed. When none is observed, but advance calculations are made with the thought in mind that one could be, then the symbol a may be thought of as a stochastic variable itself.

$f^*(a|p)$: a cumulative distribution function for the stochastic variable a described above that would exist if p were the true parameter value of the joint cumulative distribution function $F(x)$.

$f^*(a|\xi)$: the expected cumulative distribution function of a obtained by calculating the expected value of $f^*(a|p)$. That is, $f^*(a|p)$ is weighted by the a priori knowledge, ξ, of the distribution of p in Ω to obtain the average.

c: the cost of one observation

ξ_a: the a posteriori cumulative distribution function of P in Ω based upon the observation a. If ξ is an a priori distribution and a is the result of a single observation, then ξ_a is an a posteriori distribution obtained by applying Bayes theorem (Appendix A) to modify ξ to

ξ_a - the modification being based upon the observation a.

Combining these notions, it is possible to paraphrase the recursion formula as follows:

the least average risk produced by decision functions which prescribe from 0 to m + 1 observations	= the minimum of: (1) the least average risk produced by decision functions which prescribe no observation (2) the cost of one observation plus the expected value of the least average risk produced by decision functions which prescribe from 0 to m observations after the first one

This formula seems reasonable and its validity may be shown under the assumptions of Chapter III. If we want to know the least average risk to be had by allowing decision functions prescribing from 0 to m + 1 observations, we can surely get at it by breaking the decision functions we are allowing into two groups and picking the minimum one of the two least average risks attainable from these two groups. If the breakdown is made into (1) decision functions prescribing no observation and (2) decision functions prescribing from 1 to m + 1 observations, we are set up to select the minimum as indicated in the recursion formula. The least average risk attainable from the first group is simply $\rho_0(\xi)$, as previously defined. The least average risk attainable from the second group is more complicated. Since this group prescribes from 1 to m + 1 observations, we are certain to take at least one observation. This accounts for the c in the formula. After this one certain observation is taken, its value being a, it is possible to modify the a priori distribution $\xi(p)$ in Ω

to an a posteriori distribution $\xi_a(p)$ in Ω by the Bayes theorem of Appendix A. At this point we would want to proceed by using the a posteriori distribution ξ_a, since it is an improvement over the a priori distribution. To do so, we would calculate the least average risk produced by decision functions prescribing from 0 to m more observations, that is, $\rho_m(\xi_a)$ as previously defined. This would give us an expression.

$$c + \rho_m(\xi_a)$$

for the least average risk attainable from our second group of decision functions. The reasoning thus far has omitted one subtle, but key point. It is that the single observation a is never actually taken. Therefore, we must consider all possible values that a might take in a future observation. To do this we must regard the value a as a stochastic variable, and compute an expected value of $\rho_m(\xi_a)$ with respect to the distribution of a. This accounts for the fact that the second choice on the right side of the formula takes the form

$$c + \int_{-\infty}^{\infty} \rho_m(\xi_a)\, df^*(a|\xi) .$$

Wald has shown that $\rho_m(\xi)$ will, for a sufficiently large value of m, differ from $\rho(\xi)$ by an arbitrarily small amount. This permits us to write

$$\text{(B)} \qquad \lim_{m=\infty} \rho_m(\xi) = \rho(\xi) ,$$

and leads us from formula (A) to the formula

$$\text{(C)} \qquad \rho(\xi) = \text{Min} \left[\begin{matrix} \rho_o(\xi) \\ c + \int_{-\infty}^{\infty} \rho(\xi_a)\, df^*(a|\xi) \end{matrix} \right] .$$

This formula is presented in the notation of the Stieltjes Integral (see Appendix A), and does not distinguish between the case where the stochastic variable a is discrete and the case where it is continuous. If we desired to do so we could write

$$(C_1) \qquad \rho(\xi) = \text{Min} \begin{bmatrix} \rho_o(\xi) \\ c + \sum \rho(\xi_a) f^{*\prime}(a|\xi) \end{bmatrix}$$

where $f^{*\prime}$ is the bar graph of a discrete stochastic variable, and

$$(C_2) \qquad \rho(\xi) = \text{Min} \begin{bmatrix} \rho_o(\xi) \\ c + \int_{-\infty}^{\infty} \rho(\xi_a) f^{*\prime}(a|\xi)\, da \end{bmatrix}$$

where $f^{*\prime}$ is the density function of a continuous stochastic variable. The payoff of the preceding discussion lies in the manner in which $\rho(\xi)$ and $\rho_o(\xi)$ may be used to obtain a Bayes solution. It is best explained by Wald when he says:

> A Bayes solution relative to a given a priori probability measure ξ_o can immediately be given in terms of the functions $\rho(\xi)$ and $\rho_o(\xi)$ as follows: If $\rho(\xi_o) = \rho_o(\xi_o)$, do not take any observation and make a final decision d_o^t for which $W(\xi_o, d_o^t) = \rho_o(\xi_o)$. If $\rho(\xi_o) < \rho_o(\xi_o)$, take an observation on X_1 and compute the a posteriori probability measure ξ_{x_1} corresponding to ξ_o and x_1. If $\rho(\xi_{x_1}) = \rho_o(\xi_{x_1})$, stop experimentation and make a final decision d^t for which $W(\xi_{x_1}, d^t) = \rho_o(\xi_{x_1})$. If $\rho(\xi_{x_1}) < \rho_o(\xi_{x_1})$, take an observation x_2 on X_2. In general, after the observations $x_1, \ldots, x_m$ have been made, take an additional observation if $\rho(\xi_{x_1}, \ldots, x_m) < \rho_o(\xi_{x_1}, \ldots, x_m)$, and stop experimentation with a proper terminal decision if $\rho(\xi_{x_1}, \ldots, x_m) = \rho_o(\xi_{x_1}, \ldots, x_m)$, where $\xi_{x_1}, \ldots x_m$ denotes the a posteriori probability measure corresponding to

ξ_0 , x_1 , . . . , x_m .[3]

3. Stochastic Variables Limited to Two Values.

The case where the X_i are restricted to take only two values is quite special. It will arise when the value of each variable may be considered to be a failure or a success, as in Exhibit A. In such cases, the values of the stochastic variables are taken as 0 and 1 . These correspond respectively to failure and success. The following shorthand notation is used to describe cumulative distribution functions.

ξ_0 : an a priori cumulative distribution function of p in Ω

ξ_{ij} : the a posteriori distribution of p in Ω after i 0's and j 1's have been observed. ξ_{oo} is the same as ξ_0 .

If there exists a positive integer m such that

$$\rho_o(\xi_{mj}) \leq c \text{ and } \rho_o(\xi_{im}) \leq c \text{ for } i = 1,2\ldots m; \quad j = 1,2\ldots m ,$$

then it is clear from formula (C) that

$$\rho(\xi_{mj}) = \rho_o(\xi_{mj}) \text{ and } \rho(\xi_{im}) = \rho_o(\xi_{im}) \text{ for } i = 1,2\ldots m; \quad j = 1,2\ldots m .$$

This may be explained in words as follows: Suppose an integer m exists such that when either m 0's or m 1's have been observed, and the attendant a posteriori distributions computed, it is found that the least average risk attainable, by allowing, from this point on, decision functions which prescribe no experimentation, does not exceed c . Then from formula (C) the least average risk attainable by allowing decision functions prescribing any amount of exper-

[3] Wald uses the term probability measure where we have been using cumulative distribution function.

imentation is equal to that which is attainable by allowing only decision functions which call for no further experimentation.

Let us now define p_{ij} to be the probability of obtaining the value 1 on a single observation when ξ_{ij} is the a priori distribution. That is,

$$p_{ij} = \int_{\Omega} f(1|p)\, d\xi_{ij}(p),$$

is the $f^{*\prime}$ of formula (C_1).

Then the probability of obtaining the value 0 on a single trial is $1 - p_{ij}$. Using this notation, the formula (C_1) of the preceding section may be adapted to the case where the stochastic variables take only two values. It becomes

$$(D) \quad \rho(\xi_{ij}) = \operatorname{Min} \left[\begin{array}{l} \rho_o(\xi_{ij}) \\ c + p_{ij}\rho(\xi_{i,j+1}) + (1 - p_{ij})\rho(\xi_{i+1,j}) \end{array} \right].$$

It is this particular form of the formula, along with the defining relation

$$(E) \quad \rho_o(\xi_{ij}) = \operatorname*{Min}_{d^t} W(\xi_{ij}, d^t) = \operatorname*{Min}_{d^t} \int W(p, d^t)\, \xi'(p)\, dp$$

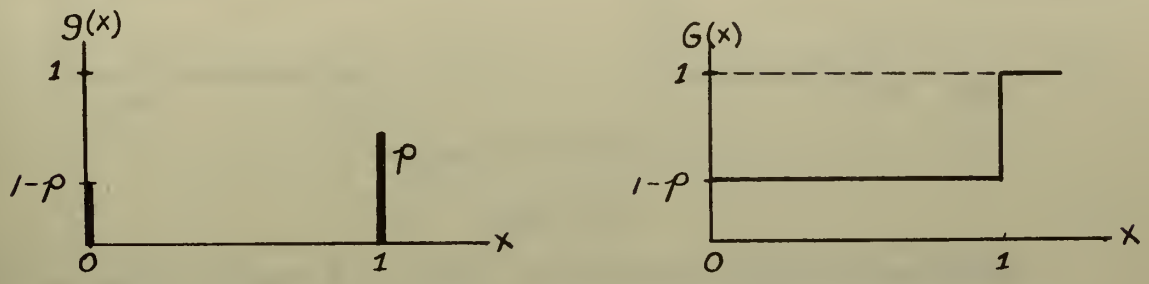
given earlier and the Bayes theorem of Appendix A that we shall use in solving Exhibit A. The details of their use are best seen by studying the detailed solution of the problem.

4. The Solution of Exhibit A.

The dollar values given in the original presentation of Exhibit A in Chapter I may be multiplied by 10^{-6} without altering the procedure followed in solving the problem. This amounts to expressing all costs in millions of dollars. Making this simple transformation and convert-

ing each original "precept" of the problem into a technical datum, as subsequently introduced, we have given:

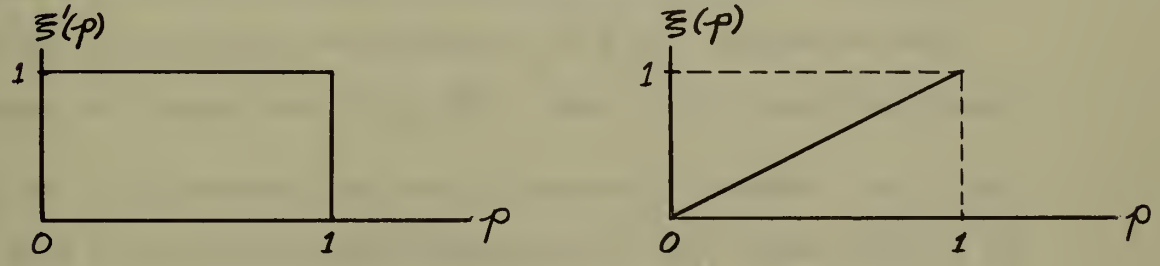
(1) the stochastic process: $X_i = 0$ (failure) or $X_i = 1$ (success)



Distribution of X

Figure 8.

(2) the a priori distribution in the parameter space:



Distribution of P

Figure 9.

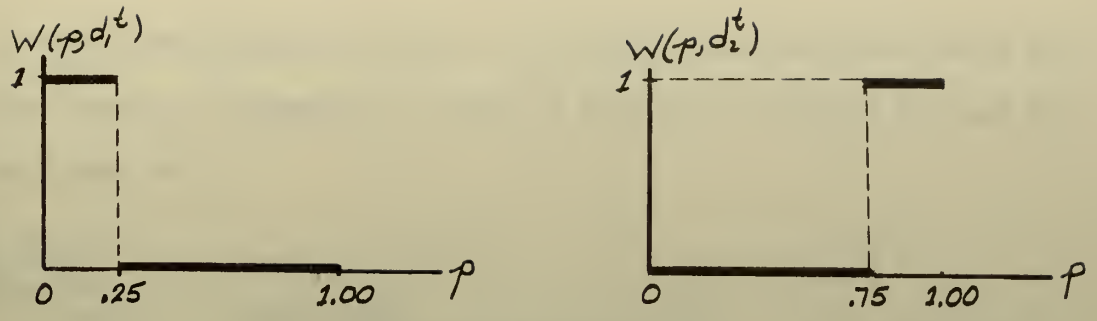
(3) the decision space: D^t consists of two elements:

d_1^t : accept the midget submarine

d_2^t : reject the midget submarine

(4) the weight function:

$$W(p, d_1^t) = 0 \text{ for } p > \frac{1}{4}$$
$$= 1 \text{ for } p < \frac{1}{4}$$
$$W(p, d_2^t) = 0 \text{ for } p < \frac{3}{4}$$
$$= 1 \text{ for } p > \frac{3}{4}$$



Weight Function

Figure 10.

(5) the cost function: $C = .004$, the cost of a single experiment.

The Bayes solution to this problem, that is, the δ_o that we seek, is a table. It is the same table that was given in Chapter I. The upper entries in the cells of the table are values of $\rho_o(\xi_{ij})$, while the lower entries are values of $\rho(\xi_{ij})$. Hence the table provides the comparison of $\rho_o(\xi_{ij})$ and $\rho(\xi_{ij})$ needed to determine how to experiment and reach a terminal decision. Our immediate task is to calculate these values of $\rho_o(\xi_{ij})$ and $\rho(\xi_{ij})$ to complete the table. We may begin by calculating the values of $\rho_o(\xi_{ij})$ for successive diagonal entries (i=j) from formula (E) and Figures 9 and 10.

$\underline{\rho_o(\xi_{oo})}$:

$$W(\xi_{oo}, d_1^t) = \int_\Omega W(p, d_1^t)\,\xi'(p)\,dp = \int_0^{1/4} (1)(1)\,dp + \int_{1/4}^{1} (0)(1)\,dp$$

$$= p\Big]_0^{1/4} = \frac{1}{4}$$

$$W(\xi_{oo}, d_2^t) = \int_\Omega W(p, d_2^t)\,\xi'(p)\,dp = \int_0^{3/4} (0)(1)\,dp + \int_{3/4}^{1} (1)(1)\,dp$$

$$= p\Big]_{3/4}^{1} = \frac{1}{4}$$

$$\rho_o(\xi_{oo}) = \min_{d^t} W(\xi_{oo}, d^t) = \min \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \end{bmatrix} = \frac{1}{4} = .2500$$

The remaining diagonal entries of $\rho_o(\xi_{ij})$ are computed using the Bayes theorem (Appendix A, Case II) as well as formula (E) and Figures 9 and 10.

$\rho_o(\xi_{11})$:

$$\xi'_{11} = \frac{(1)(1-p)(p)}{\int_o^1 (1)(1-p)(p)dp} = \frac{p - p^2}{\frac{p^2}{2} - \frac{p^3}{3}}\Big]_o^1 = 6p - 6p^2$$

$$W(\xi_{11}, d_1^t) = \int_o^{1/4} (1)(6p - 6p^2)\,dp + \int_{1/4}^1 (0)(6p - 6p^2)\,dp$$

$$= 3p^2 - 2p^3\Big]_o^{1/4} = \frac{5}{32}$$

$$W(\xi_{11}, d_2^t) = \int_o^{3/4} (0)(6p - 6p^2)\,dp + \int_{3/4}^1 (1)(6p - 6p^2)\,dp$$

$$= 3p^2 - 2p^3\Big]_{3/4}^1 = \frac{5}{32}$$

$$\rho_o(\xi_{11}) = \min_{d^t} W(\xi_{11}, d^t) = \text{Min}\left[\frac{5}{32}, \frac{5}{32}\right] = \frac{5}{32} = .1563$$

The procedure may be generalized for all diagonal entries $(i = j)$ so that we have

$$\rho_o(\xi_{ii}) = \int_o^{1/4} \frac{(1-p)^i (p)^i}{\int_o^1 (1-p)^i (p)^i\,dp}\,dp = \frac{\int_o^{1/4} (1-p)^i (p)^i\,dp}{\int_o^1 (1-p)^i (p)^i\,dp}$$

for all i. Values of this last expression may be obtained from Tables of the Incomplete Beta Function.[4] The use of these tables permits easy evaluation. Values obtained are the entries shown in the upper halves of the diagonal cells in Table 3.

The next step is to calculate the non-diagonal $(i \neq j)$ upper entries. This is done as follows:

$\rho_o(\xi_{23})$:

[4] Tables of the Incomplete Beta Function, Pearson, University Press, Cambridge, 1934.

$$\xi'_{23} = \frac{(1)(1-p)^2 p^3}{\int_0^1 (1)(1-p)^2 p^3 dp} = \frac{p^3 - 2p^4 + p^5}{\left[\frac{p^4}{4} - \frac{2p^5}{5} + \frac{p^6}{6}\right]_0^1} = 60\,(p^3 - 2p^4 + p^5)$$

$$W(\xi_{23}, d_1^t) = \int_0^{1/4} (1)[60(p^3 - 2p^4 + p^5)]\,dp = \left[15p^4 - 24p^5 + 10p^6\right]_0^{1/4}$$

$$= .0376$$

$$W(\xi_{23}, d_2^t) = \int_{3/4}^1 (1)[60(p^3 - 2p^4 + p^5)]\,dp = \left[15p^4 - 24p^5 + 10p^6\right]_{3/4}^1$$

$$= .1700$$

$$\rho_o(\xi_{23}) = \text{Min}\begin{bmatrix} .0376 \\ .1700 \end{bmatrix} = .0376$$

Again the procedure generalizes and we have, for $i < j$,

$$\rho_o(\xi_{ij}) = \frac{\int_0^{1/4} (1-p)^i (p)^j\,dp}{\int_0^1 (1-p)^i (p)^j\,dp} .$$

For $i > j$ we have

$$\rho_o(\xi_{ij}) = \frac{\int_{3/4}^1 (1-p)^i (p)^j\,dp}{\int_0^1 (1-p)^i (p)^j\,dp} .$$

As before, the evaluation may be accomplished by use of the Tables of the Incomplete Beta Function. Note that $\rho_o(\xi_{ij}) = \rho_o(\xi_{ji})$. This makes it necessary to evaluate entries on only one side of the main diagonal, since the remaining entries may be determined by symmetry.

With the upper entries filled in, we turn our attention to the lower entries. They may be determined in two stages. The first stage is just to compare $\rho_o(\xi_{ij})$ with c for each square. Since

$$(D)\quad \rho(\xi_{ij}) = \operatorname{Min}\begin{bmatrix} \rho_o(\xi_{ij}) \\ c + p_{ij}\,\rho(\xi_{i,j+1}) + (1 - p_{ij})\,\rho(\xi_{i+1,j}) \end{bmatrix},$$

we may immediately select $\rho_o(\xi_{ij})$ as the value of $\rho(\xi_{ij})$ for all squares in which $\rho_o(\xi_{ij}) \leq c$. For those squares in which $\rho_o(\xi_{ij}) > c$ we must use formula (D) and the formula for p_{ij} to calculate $\rho(\xi_{ij})$. For example, in the case of diagonal entries where $p_{ii} = \frac{1}{2}$, we may compute

$$\rho(\xi_{99}) = \operatorname{Min}\begin{bmatrix} \rho_o(\xi_{99}) \\ c + p_{99}\,\rho(\xi_{9,10}) + (1 - p_{99})\,\rho(\xi_{10,9}) \end{bmatrix}$$

$$= \operatorname{Min}\begin{bmatrix} .0090 \\ .004 + \frac{1}{2}(.0039) + \frac{1}{2}(.0039) \end{bmatrix} = \operatorname{Min}\begin{bmatrix} .0090 \\ .0079 \end{bmatrix}$$

$$= .0079$$

In the case of non-diagonal entries, the first step is to compute p_{ij} from the formula

$$p_{ij} = \int_{\Omega} f(1|p)\, d\xi_{ij} = \int_{\Omega} f(1|p)\, \xi'_{ij}(p)\, dp .$$

Upon substituting in this formula we have

$$p_{ij} = \int_0^1 (p)\, \frac{(1-p)^i (p)^j}{\int_0^1 (1-p)^i (p)^j\, dp}\, dp = \frac{\int_0^1 (1-p)^i (p)^{j+1}\, dp}{\int_0^1 (1-p)^i (p)^j\, dp} .$$

This last expression may be evaluated using the Tables of the Incomplete Beta Function. Once p_{ij} is known, we have only to solve formula (D) for $\rho(\xi_{ij})$. Values of $\rho(\xi_{ij})$ obtained in this way complete Table 3.

The best sequence for calculating the lower entries is as follows: Fill in the main diagonal entry in the lower right hand corner first. Then progress to the left in that row. Next, move up to the next higher diagonal entry and again work left on the row. The entries on the upper right hand side of the diagonal can be filled in by symmetry.

The interpretation of the table, as given in Chapter I, may now be stated in terms of the technical notation. Begin taking observations and after each observation compare $\rho(\xi_{ij})$ with $\rho_o(\xi_{ij})$. As long as $\rho(\xi_{ij})$ is less than $\rho_o(\xi_{ij})$, continue taking observations. When an observation is made such that $\rho(\xi_{ij}) = \rho_o(\xi_{ij})$, stop experimentation and make a proper terminal decision. If $i > j$, the terminal decision will be to reject the midget submarine. If $i < j$, the terminal decision will be to accept the midget submarine.

Number of 1's

Number of 0's	0	1	2	3	4	5	6	7	8	9	10
0	.2500 .0252	.0625 .0212	.0156 .0126	.0039 .0039	.0010 .0010	.0002 .0002	.0001 .0001	.0000 .0000	.0000 .0000	.0000 .0000	.0000 .0000
1	.0625 .0212	.1563 .0265	.0508 .0225	.0156 .0141	.0046 .0046	.0013 .0013	.0004 .0004	.0001 .0001	.0000 .0000	.0000 .0000	.0000 .0000
2	.0156 .0126	.0508 .0225	.1035 .0250	.0376 .0210	.0129 .0129	.0042 .0042	.0013 .0013	.0005 .0005	.0001 .0001	.0000 .0000	.0000 .0000
3	.0039 .0039	.0156 .0141	.0376 .0210	.0706 .0225	.0273 .0185	.0100 .0100	.0035 .0035	.0012 .0012	.0004 .0004	.0001 .0001	.0000 .0000
4	.0010 .0010	.0046 .0046	.0129 .0129	.0273 .0185	.0489 .0210	.0198 .0161	.0076 .0076	.0028 .0028	.0010 .0010	.0003 .0003	.0001 .0001
5	.0002 .0002	.0013 .0013	.0042 .0042	.0100 .0100	.0198 .0161	.0343 .0176	.0143 .0136	.0056 .0056	.0022 .0022	.0008 .0008	.0003 .0003
6	.0001 .0001	.0004 .0004	.0013 .0013	.0035 .0035	.0076 .0076	.0143 .0136	.0243 .0143	.0103 .0103	.0042 .0042	.0016 .0016	.0006 .0006
7	.0000 .0000	.0001 .0001	.0005 .0005	.0012 .0012	.0028 .0028	.0056 .0056	.0103 .0103	.0173 .0115	.0075 .0075	.0031 .0031	.0012 .0012
8	.0000 .0000	.0000 .0000	.0001 .0001	.0004 .0004	.0010 .0010	.0022 .0022	.0042 .0042	.0075 .0075	.0124 .0094	.0054 .0054	.0023 .0023
9	.0000 .0000	.0000 .0000	.0000 .0000	.0001 .0001	.0003 .0003	.0008 .0008	.0016 .0016	.0031 .0031	.0054 .0054	.0090 .0079	.0039 .0039
10	.0000 .0000	.0000 .0000	.0000 .0000	.0000 .0000	.0001 .0001	.0003 .0003	.0006 .0006	.0012 .0012	.0023 .0023	.0039 .0039	.0064 .0069

Solution of Exhibit A in Technical Form

Table 3.

CHAPTER V

THE MINIMAX SOLUTION

1. The Minimax Solution and its Relation to the Bayes Solution.

The scope of this paper, for detailed discussion, is limited to the Bayes solution of the statistical decision problem. Emphasis is given to the special case in which the X_i are independently and identically distributed, and confined to take only two values. However, to avoid having the reader assume that this constitutes all of statistical decision theory, mention should be made of the Minimax solution.

It was pointed out in Chapter II that the Bayes solution is always given relative to an a priori distribution of the unknown parameter. If such a distribution cannot be given, it may still be possible to solve a statistical decision problem. A solution may be obtained by viewing the decision problem as a zero sum, two person game, and solving the game. A solution obtained in this manner is termed a Minimax solution. A Minimax solution may also be obtained in other ways. A Minimax solution, as noted in Chapter III, is a particular Bayes solution. Specifically, it is that Bayes solution which is given relative to the least favorable a priori distribution of the unknown parameter.

The difference between the Bayes solution and the Minimax solution lies in the choice of a yardstick for comparing the relative merits of the various decision functions. The basic criterion, in either case, is the risk function. But the modification of this criterion, to arrive at the final yardstick, is different. The reader will

recall from Chapter II that, for a Bayes solution, the risk function, $r(p,\delta)$, was modified to an expected risk, $r(\xi,\delta)$, by averaging out the p, and this expected risk constituted the final yardstick. The modification was accomplished by using the a priori distribution, $\xi(p)$. The expected risk, which could then be ordered as to magnitude where the magnitude depended on δ alone, permitted the selection of the particular statistical decision function, δ_o, that provided the least expected risk and hence the optimum solution relative to the assumed $\xi(p)$. In the case of a Minimax solution, we consider that an a priori distribution is not available. Hence, the procedure employed to modify the risk function to a suitable final yardstick must be altered. The procedure that is used consists of taking the maximum risk vice the expected risk. An a priori distribution of p is not required to do this. We simply take the maximum value of the risk, $r(p,\delta)$, for each δ, by selecting the p that maximizes it. That is,

$$\text{Max risk} = \max_{p \in \Omega} r(p,\delta) .$$

This new function, the maximum risk, is dependent on δ alone, and can therefore be ordered as to magnitude with the magnitude determined by δ. Again we select the particular δ_o that minimizes the yardstick. That is, we take

$$\min_{\delta} \max_{p} r(p,\delta) \quad \text{for all } \delta.$$

This is sometimes written

$$\max_{p} r(p,\delta_o) \leq \max_{p} r(p,\delta) \quad \text{for all } \delta.$$

The δ_o of this latter expression constitutes a Minimax solution.

The statement that the Minimax solution is a Bayes solution

relative to the least favorable a priori distribution now seems reasonable. For, if our a priori distribution for a Bayes solution were the least favorable of all, it would lead us to the $\underset{p}{\text{Max}}\ r(p, \delta)$ as a yardstick.

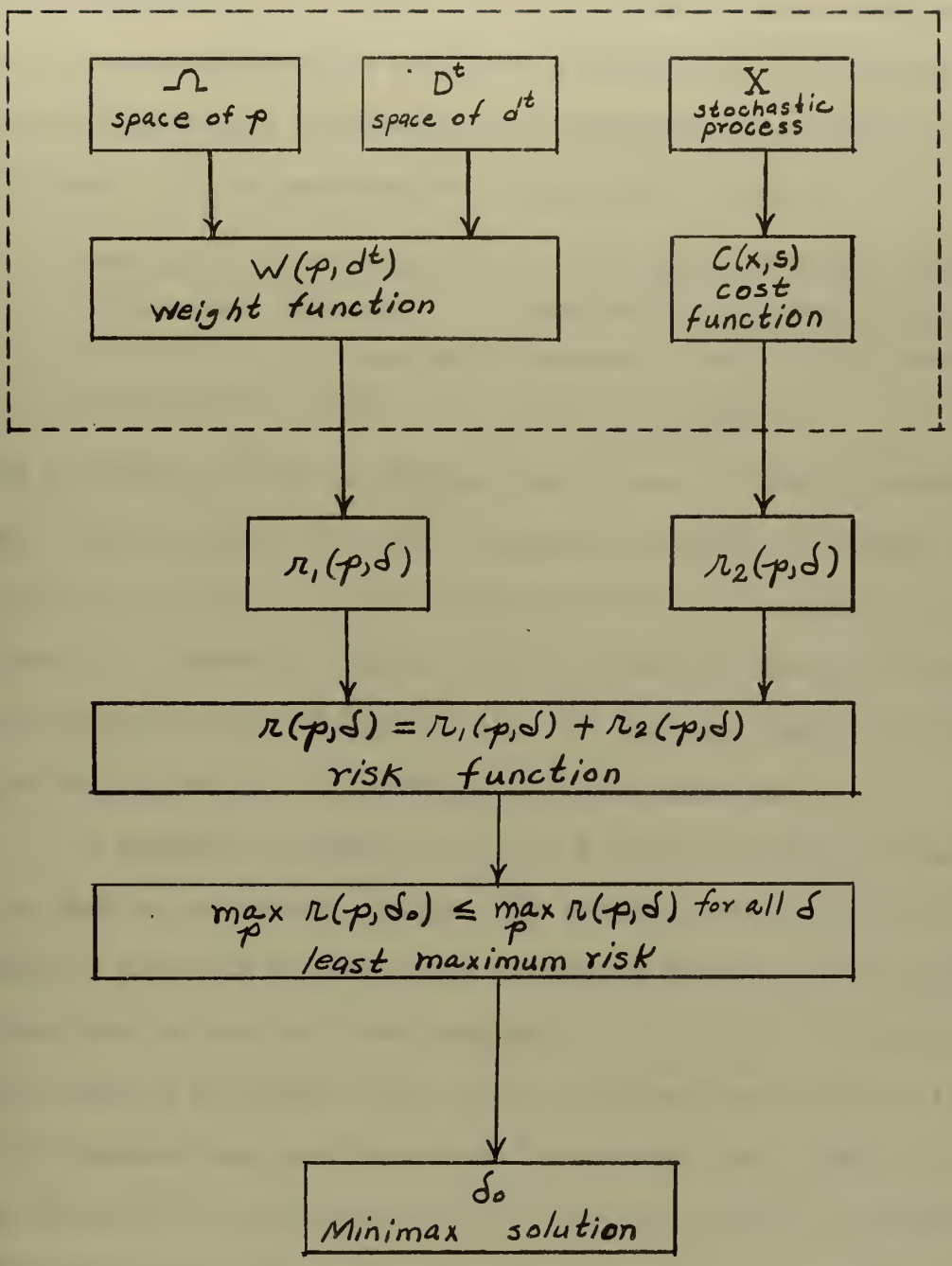
The procedure used to arrive at a Minimax solution is summarized in Figure 11 (see following page).

2. Relation to the Theory of Games.

The reader familiar with von Neumann's theory of games will recognize the procedure of the preceding section as essentially the same as that of game theory. In fact, Wald points out the detailed correspondence between the statistical decision problem in which no a priori distribution is given and the zero sum, two person game. In the general case, the corresponding game is a continuous one. This means that the question of the strict determinateness of the game must be investigated. Whereas the fundamental theorem of rectangular games assures the existence of a solution to any finite game, no such assurance exists in the case of all infinite games. However, Wald demonstrates that, under suitable assumptions, any statistical decision problem viewed as a continuous game may be approximated arbitrarily closely by a finite game. This means that, even if the continuous game is not strictly determined, no practical limitation is imposed. The detailed procedure employed in arriving at a Minimax solution of a statistical decision problem in this manner involves the formulation of the problem as a game, and the solution of the game. It will not be covered here.

3. Summary.

When an operations analyst is confronted with the need to make a decision on the basis of the results of conducted trials, the accuracy



Block Diagram of the Buildup to a Minimax Solution

Figure 11 .

of which depends upon the true value of an unknown parameter, and the cost of the experimentation required to estimate the value of the parameter is significant, a statistical decision problem is indicated. As Wald puts it, in two sentences here taken out of context,

> A statistical decision problem is formulated with reference to a stochastic process . . . A statistical decision problem with reference to a stochastic process X arises only when the distribution F (x) is not completely known.

Once a statistical decision problem has arisen, it must be possible to specify the stochastic process, the parameter space, the space of terminal decisions, the weight function and the cost function, in order to solve it. A solution consists of determining the particular statistical decision function that prescribes the optimum plan for conducting experimentation and reaching a terminal decision.

The procedure employed to reach a solution involves the use of a risk function as a basic criterion for selection the optimum decision function. This risk function takes account of both the cost of wrong decision and the cost of experimentation. If an a priori distribution of the unknown parameter can be given, the final yardstick for selecting the optimum decision function is the average risk; if not, the final yardstick is the maximum risk. In either case, the yardstick is ordered as to magnitude, and that decision function which provides the least value of the yardstick is selected as a solution. The first case yields a Bayes solution; the second a Minimax solution.

The consequence of the final decision is probabilistic. This

of which depends upon the true value of an unknown parameter, and the cost of the experimentation required to estimate the value of the parameter is significant, a statistical decision problem is indicated. As Wald puts it, in two sentences here taken out of context,

> A statistical decision problem is formulated with reference to a stochastic process . . . A statistical decision problem with reference to a stochastic process X arises only when the distribution F (x) is not completely known.

Once a statistical decision problem has arisen, it must be possible to specify the stochastic process, the parameter space, the space of terminal decisions, the weight function and the cost function, in order to solve it. A solution consists of determining the particular statistical decision function that prescribes the optimum plan for conducting experimentation and reaching a terminal decision.

The procedure employed to reach a solution involves the use of a risk function as a basic criterion for selection the optimum decision function. This risk function takes account of both the cost of wrong decision and the cost of experimentation. If an a priori distribution of the unknown parameter can be given, the final yardstick for selecting the optimum decision function is the average risk; if not, the final yardstick is the maximum risk. In either case, the yardstick is ordered as to magnitude, and that decision function which provides the least value of the yardstick is selected as a solution. The first case yields a Bayes solution; the second a Minimax solution.

The consequence of the final decision is probabilistic. This

means that the final decision may, in a particular instance, conceivably be a poor one. Nonetheless, the theory offers a rational approach to the type of problem it fits, and is superior to any other known approach.

BIBLIOGRAPHY

1. Wald, A. — STATISTICAL DECISION FUNCTIONS. John Wiley and Sons, Inc. New York, 1950

2. Blackwell, D. and Girshick, M. A. — THEORY OF GAMES AND STATISTICAL DECISIONS. John Wiley and Sons, Inc. New York, 1954

3. Pearson, K. — TABLES OF THE INCOMPLETE BETA-FUNCTION. The University Press Cambridge, 1934

APPENDIX A

SOME SELECTED MATHEMATICAL CONCEPTS

1. Probability.

Probability is a quantitative measure of the likelihood of the occurrence of events. It is expressed by assigning a number in the range $(0, 1)$ to any specific event. For example, if an event is certain to occur it has probability 1; if it is certain not to occur it has probability 0. If an event has a fifty-fifty chance of occurring it has probability $\frac{1}{2}$. The probability of an event may be estimated by conducting repeated trials and employing the formula

$$\text{probability} = \frac{\text{number of successes}}{\text{number of trials}} .$$

2. Stochastic Variables.

A stochastic variable may be defined to be a function which associates a real number with every possible outcome of an experiment. The outcome of any particular performance of the experiment is said to be a value assumed by the stochastic variable, it being understood that this outcome is a chance occurrence. A stochastic variable is termed discrete if the number of distinct values which it may assume is either finite or may be arranged in a sequence (i.e., is denumerable). It is termed continuous if its possible values may be represented by an interval on the real line, e.g., all the points x such that $a < x < b$ or $-\infty < x < \infty$.

3. The Distribution of a Discrete Stochastic Variable.

The correspondence between the values of a discrete stochastic variable

and the probabilities that it will take on these values may be described either by a probability function (bar graph) or by a cumulative probability distribution function (step function). As an example of this, consider a single true die to be tossed a large number of times. A mathematical description of the stochastic nature of this experiment may be formulated as follows:

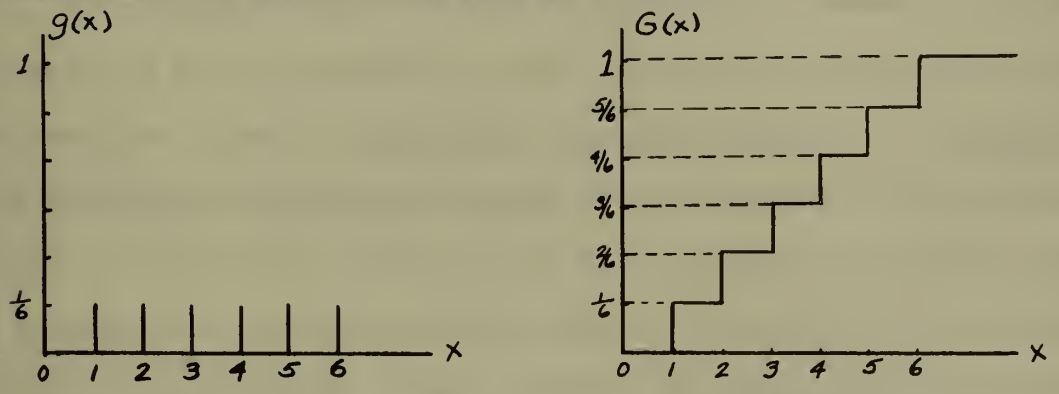
X: a stochastic variable representing the value shown on the die after any throw.

x_i: real values which may be assumed by the stochastic variable X, i.e., 1, 2, 3, 4, 5, and 6.

G(x): the probability that X will take on a value less than or equal to x. $G(x) = Pr(X \leq x)$.

g(x): the probability that X will take on the value x, $g(x) = Pr(X = x)$.

These quantities may be displayed as follows:



Bar Graph Step Function

Distribution of A Discrete Stochastic Variable

Figure 12.

The bar graph indicates that the probability of tossing any particular number on a given throw is the same for all numbers, and is equal to $\frac{1}{6}$. The step function is an alternative way of presenting essentially

the same information. It permits the probability that a toss will show a value less than or equal to any given value to be read directly. For instance, the probability that the die will show three of less on a throw is

$$G(3) = \frac{3}{6} = \frac{1}{2} ,$$

a result that would be anticipated. It is to be noted that

$$\sum_{i=1}^{6} g(x_i) = 1 \quad \text{and} \quad g(x_i) \geq 0 \qquad i = 1,2,3,4,5,6.$$

Also,

$$G(0) = 0 \quad \text{and} \quad G(6) = 1 .$$

These are fundamental relations associated with the probability function and cumulative probability distribution function of the stochastic variable X .

4. The Distribution of a Continuous Stochastic Variable.

The correspondence between the values of a continuous stochastic variable and the probabilities that it will take on these values may be described either by a probability density function or by a cumulative probability distribution function. As an example of this, consider a line six units long on which a point is to be chosen at random. This is an experiment similar to the one used to describe the distribution of a discrete stochastic variable, but now the value of the stochastic variable may be any number in the closed interval $[0,6]$. A mathematical description of the stochastic nature of this experiment may be formulated as follows:

X: a stochastic variable representing the coordinate point selected on any try

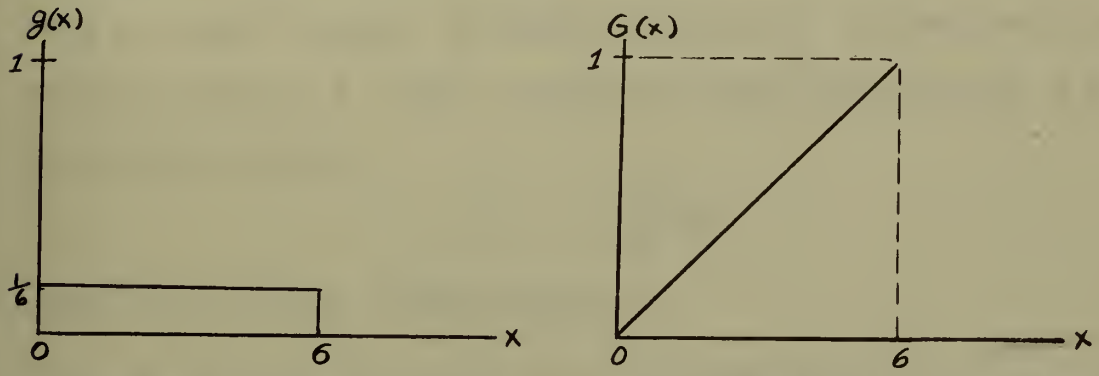
x: real values which the stochastic variable X may assume.

$G(x)$: the probability that X will take on a value less than or equal to x. $G(x) = \Pr(X \leq x)$.

$g(x)$: the probability density function of X.

$g(x)\,dx$: the probability that X will take on a value between x and $x + dx$. $g(x)\,dx = \Pr(x \leq X \leq x + dx)$.

The probability density function and the cumulative probability distribution function associated with X may be displayed as follows:



Probability Density Function Cumulative Distribution Function

Distribution of a Continuous Stochastic Variable

Figure 13.

The particular density function of Figure 13. is said to be uniform. This means that the stochastic variable is equally likely to take on any one of its values and accounts for the straight, horizontal line which represents the density function. Other stochastic variables may have a bias such that some of the values are more likely to occur than others, and will have density functions which are not represented by horizontal lines. In any case, the area under the density function will always be 1, and the cumulative distribution function will increase monotonically to a maximum value of 1 for increasing values of x. It is to be noted that

$$\int_0^6 g(x)\,dx = 1 \quad \text{and} \quad g(x) \geq 0 \quad \text{for} \quad 0 \leq x \leq 6 .$$

Also,

$$G(0) = 0 \qquad \text{and} \qquad G(6) = 1.$$

These are fundamental relations associated with the probability density function and the cumulative probability distribution function of the stochastic variable X. In texts on probability theory, it is shown that, for continuous stochastic variables, the density function, when it exists, is the derivative of the cumulative distribution function. This is a basic relation. It should be noted that, whereas every stochastic variable, X, has a cumulative distribution function $G(x)$, the density function

$$g(x) = \frac{d\,G(x)}{dx}$$

exists only if $G(x)$ is differentiable.

5. The Expected Value of a Stochastic Variable.

The expected value (average value) of a discrete stochastic variable is defined to be

$$(5.1) \qquad E(X) = \bar{x} = \sum_{all\ i} x_i\, g(x_i) ,$$

and the expected value of a continuous stochastic variable which has a density function is defined to be

$$(5.2) \qquad E(X) = \bar{x} = \int_{all\ x} x\, g(x)\, dx .$$

The expectation, $E(X)$, is often termed a weighted average. In the case of a discrete stochastic variable, the "weight" associated with x_i is $g(x_i)$. In the case of a continuous stochastic variable, the "weight" associated with x is $g(x)\,dx$, i.e., the probability that x lies in dx. In the latter case we say, more briefly, that x is weighted by $g(x)$, the value of the probability density function.

In theoretical discussions it may not be desirable to distinguish between discrete and continuous stochastic variables nor to emphasize continuous stochastic variables which have a density function. In such circumstances, generally, it is only necessary to refer to the stochastic variable, say X, and its cumulative probability distribution function, say $G(x)$. To represent the expected value of X, it is customary to write

$$(5.3) \qquad E(X) = \int_{-\infty}^{\infty} x\,dG(x) \text{ , where}$$

the integral on the right represents either a Riemann-Stieltjes or a Lebesgue-Stieltjes integral according to the degree of generality of the theory of integration under consideration. In this paper, such integrals will be viewed as Riemann-Stieltjes integrals. In short, we may regard the integral of equation (5.3) as a concept which includes both of the concepts of equations (5.1) and (5.2) as special cases. We should note that the expected value of a function of a stochastic variable may also be defined. For example, if $h(X)$ is a function of the stochastic variable X, we may write

$$E\left[h(X)\right] = \int_{-\infty}^{\infty} h(x)\,dG(x) .$$

As a further illustration of the notion of expectation and the use of general functional notation, consider the following:

P: a stochastic variable

p: real values that may be assumed by P

$\xi(p)$: cumulative probability distribution function of P

$\xi'(p)$: probability density function of P .

The expected value of P , according to equation (5.2) , is

$$E(P) = \bar{p} = \int_{\text{all } p} p \; \Xi'(p)\, dp \; .$$

This may be more conveniently expressed, using the Riemann-Stieltjes integral, as

$$\bar{p} = \int_{\Omega} p \, d\Xi \; ,$$

where Ω is the space of p , and the differential $d\Xi$ is used instead of $\Xi' dp$. Thus, values of p are weighted by $d\Xi$, where formerly values of p were weighted by $\Xi'(p)\, dp$. The symbol

$$\int p \, d\Xi$$

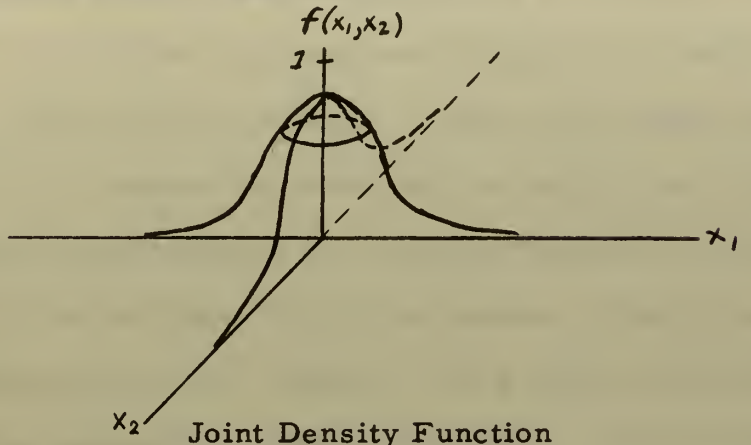
is a functional symbol used to express the notion of the weighting and summing. When actual computations are carried out, $d\Xi(p)$ is replaced by its equivalent expression in p , and the integration is carried out just as in elementary calculus.

6. Joint Distribution Functions.

So far we have discussed only distribution functions of a single stochastic variable X . The notion of joint distribution functions of more than one stochastic variable is often employed in probability theory. This is nothing more than an extension of the idea of the distribution function of a single stochastic variable. For example, if X_1 and X_2 are two stochastic variables, then their joint cumulative distribution function, $F(x_1, x_2)$, is the probability that $X_1 \leq x_1$ and $X_2 \leq x_2$ simultaneously. That is

$$F(x_1, x_2) = \Pr(X_1 \leq x_1 \text{ and } X_2 \leq x_2) \text{ simultaneously.}$$

The density function of such a distribution may be represented as a surface in three dimensional space as follows:



Joint Density Function

Figure 14.

It is to be noted that

$$\iint f(x_1, x_2)\, dx_1\, dx_2 = 1 \quad \text{and} \quad f(x_1, x_2) \geq 0 \quad \text{for all} \quad x_1, x_2$$

Also,

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1\, \partial x_2} .$$

These are fundamental relations associated with the joint distribution. In a similar manner, the analytical notion of a joint distribution function may be extended to any number of stochastic variables, although the geometrical representation does not apply for more than two.

7. Bayes Theorem.

Perhaps the single mathematical concept most vital to an understanding of statistical decision theory is the Bayes theorem of inverse probability. To explain this theorem in terms of the example of Chapter I, let us recall that we assumed that P, the true percentage success of the midget submarine in a future war, is equally

likely to have any value between 0 and 100%. This is equivalent to assuming that the a priori distribution of the parameter is uniform and, at the outset, represents our best knowledge of P. As the problem progresses, observations are made. These observations add to our knowledge of P, and we therefore wish to modify the originally assumed a priori distribution of P to what we term an a posteriori distribution of P on the basis of the observations. Bayes theorem provides the means to do this. That is, if an a priori distribution is known and observations are subsequently made, Bayes theorem may be used to modify the a priori distribution to an a posteriori distribution on the basis of the observations. Two forms of Bayes theorem in its application to density functions in statistical decision theory are:

Case I (Ω Discrete):

$$m\xi'_j(p) = \frac{\xi'_j(p) \prod_{i=1}^{m} g(x_i|j)}{\sum_{j=1}^{n} \xi'_j(p) \prod_{i=1}^{m} g(x_i|j)}$$

Case II (Ω Continuous):

$$m\xi'(p) = \frac{\xi'(p) \prod_{i=1}^{m} g(x_i|p)}{\int_{\Omega} \xi'(p) \prod_{i=1}^{m} g(x_i|p)\, dp}$$

In these formulas, $g(x_i|j)$ is a probability function when X is discrete and a probability density function when X is continuous, and the integer n is the number of values P may assume in Case I.

To examine the theorem further, let us study an example which finds application in this paper. Suppose

X_i $(i=1,2,\ldots)$: a collection of independently and identically distributed discrete stochastic variables

x_i $(i=1,2,\ldots)$: real values that may be assumed by each X_i . Each x_i is confined to the two values 0 or 1 .

$G(x)$: the common cumulative probability distribution (step) function applicable to each of the X_i .

$g(x)$: the common probability function (bar graph) applicable to each of the X_i .

P: a continuous stochastic variable representing the parameter of $G(x)$ or $g(x)$.

p: real values that may be assumed by P , i.e. $0 \leq p \leq 1$.

$\xi(p)$: the a priori cumulative probability distribution function of P .

$\xi'(p)$: the a priori probability density function of P .

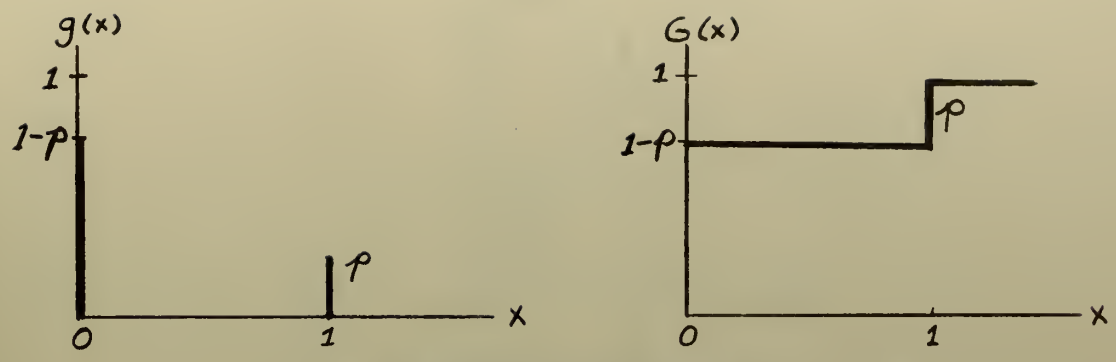
${}_m\xi(p)$: the a posteriori cumulative distribution function of P after m observations on the X_i .

${}_m\xi'(p)$: the a posteriori density function of P , after m observations on the X_i .

The Bayes formula for the density functions of P is, as in Case II.,

$$ {}_m\xi'(p) = \frac{\xi'(p) \prod_{i=1}^{m} g(x_i|p)}{\int_{\Omega} \xi'(p) \prod_{i=1}^{m} g(x_i|p)\, dp} , $$

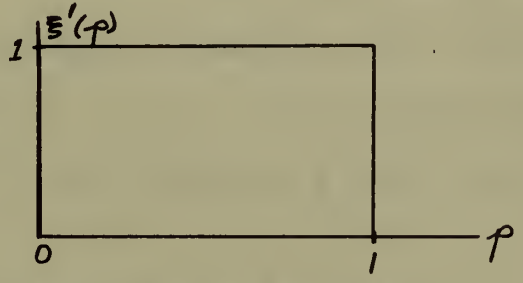
where Ω is the space of P . Let us amplify this with some diagrams and sample computations. Suppose each X_i is distributed according to the following diagram:



Distribution of the X

Figure 15.

The letter p stands for a value of the unknown parameter. If we assume the a priori density function of P to be uniform, it may be pictured as follows:



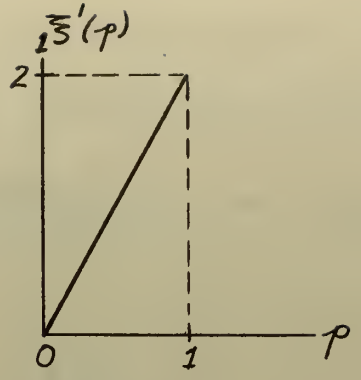
A Priori Density Function of P

Figure 16.

If we take a single observation on one of the X_i with the result $x_1 = 1$, we may apply Bayes formula as follows:

$$ {}_1\xi'(p) = \frac{(1)(p)}{\int_0^1 (1)\, p\, dp} = 2p . $$

This a posteriori density function may be pictures as follows:



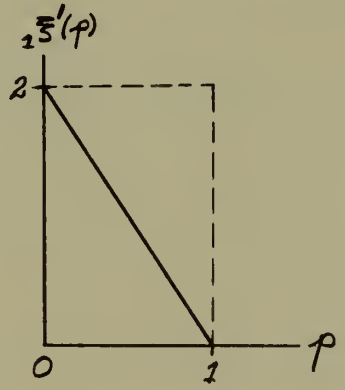
A Posteriori Density Function of P for $x_1 = 1$.

Figure 17 .

Note that the result of the single observation, through the Bayes formula, has modified the density function of P from uniform to a bias in favor of the value 1 . This is an intuitively reasonable result, since the value $x_1 = 1$ was observed. Similarly, if the result of the single observation had been $x_1 = 0$, the a posteriori density function would have been modified from uniform to a bias in favor of the value 0. In that case,

$$_1\xi'(p) = \frac{(1)(1-p)}{\int_0^1 (1)(1-p)dp} = 2 - 2p ,$$

and the a posteriori density function becomes the following:



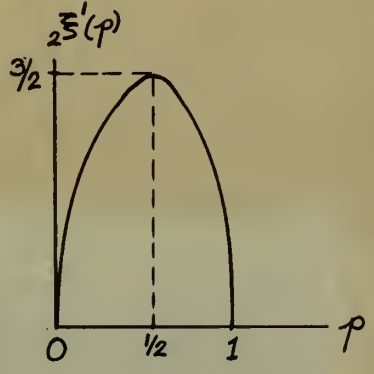
A Posteriori Density Function of P for $x_1 = 0$

Figure 18.

Again, if two observations had been taken with the results $x_1 = 1$ and $x_2 = 0$, then the a posteriori density function would be

$$ {}_2\xi'(p) = \frac{(1)(p)(1-p)}{\int_0^1 (1)(p)(1-p)\,dp} = 6p - 6p^2 , $$

and is pictured as follows:



A Posteriori Density Function of P for $x_1 = 1$, $x_2 = 0$.

Figure 19.

Note that this last density function is a parabola, and has been modified from uniform to a bias in favor of the value $\frac{1}{2}$. This is again an intuitively reasonable result to follow from the two observations.